Stored Value Cards: Costly Private Substitutes for Government Currency

Jeffrey M. Lacker

Stored value cards look like credit cards but are capable of storing monetary value. There are a number of stored value card systems being developed in the United States, and some have already been implemented in Europe and elsewhere. Stored value cards are particularly well-suited for transactions that would otherwise be carried out with currency and thus are a private substitute for government fiat money, like private bank notes. Unlike bank notes, however, stored value cards employ new technologies that are quite different from, and potentially more costly than, the coins and paper currency they are aimed at replacing. This article explores the basic welfare economics of costly private substitutes for government currency, an important class of payments system innovations.[1]

Consumers and merchants are likely to benefit from the introduction of stored value cards. Many might prefer to avoid the inconvenience and cost of handling paper currency. The usual presumption, in the absence of market imperfection, is that a successful new product must provide social benefits in excess of social costs. Issuers will attempt to cover their costs and earn a competitive return while providing consumers and merchants with a means of payment they prefer. They can do so if consumers and merchants are collectively willing to pay enough, either directly or indirectly, to remunerate

The author would like to thank Michelle Kezar and Helen Upton for research assistance, reviewers Mike Dotsey, Tom Humphrey, Ned Prescott, and John Weinberg for extensive comments, James McAfee for useful references, Urs Birchler, Scott Freeman, Marvin Goodfriend, Don Hester, Dave Humphrey, Peter Ireland, Bob King, Bennett McCallum, Geoffrey Miller, Kevin Reffett, Will Roberds, Stacey Schreft, and Bruce Summers for helpful comments and conversations, and an anonymous reviewer for helpful comments. The author is solely responsible for the contents of this article. The views expressed do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] I will use the term "currency" rather than "currency and coin" to refer to government-supplied notes and coinage; the analysis applies equally to government-minted coin.

issuers for the opportunity cost of the inputs devoted to the alternative means of payment. If stored value thrives, standard reasoning suggests that it must be because the value to consumers and merchants exceeds the cost of provision.

The fact that stored value is a monetary asset provides further reason to believe that it will be beneficial. Currency is subject to an implicit tax due to inflation, which reduces its rate of return relative to other risk-free nominal assets. Like any other (non-lump-sum) tax, the inflation tax distorts economic decisions, giving rise to deadweight costs as people try to economize on the use of currency. Private substitutes for currency provide a means of avoiding the seigniorage tax, alleviating the deadweight loss associated with any given inflation rate. Stored value cards can increase economic welfare by easing the burden of inflation.

Stored value cards could be socially wasteful, however. Stored value liabilities compete with an asset, currency, that pays no interest, while issuers are free to invest in interest-earning assets. Thus one portion of the return to stored value issuers is the spread between market interest rates and the rate of return on currency. This return far exceeds the government's cost of producing and maintaining the supply of currency—less than two-tenths of a cent per year per dollar of currency outstanding.[2] At current interest rates the private incentive to provide stored value exceeds the social cost of the currency replaced by as much as 4 or 5 cents per dollar. Thus stored value cards, if successful, will replace virtually costless government currency with a substitute that could cost substantially more.

This article presents a model in which both currency and stored value are used to make payments. Stored value cards are provided by a competitive intermediary sector and are used in transactions for which the cost of stored value is less than the float cost associated with using currency. Conditions are identified under which the equilibrium allocation of the economy with stored value cards does or does not Pareto-dominate that of an otherwise identical economy without stored value cards. The critical condition is a boundary on the average cost of stored value: stored value is beneficial or harmful depending upon whether, other things equal, average cost is below or above a certain cutoff. If average cost is low, the reduction in the deadweight loss due to inflation will be large and the resource cost will be low. If average cost is high, the resource diversion will be large and there will be little effect on the burden of inflation.

The fact that costly private substitutes for government fiat money can reduce welfare was demonstrated by Schreft (1992a), and the model presented below is an extension of hers. This fact should not be surprising—as I argue below, we should expect the same result in any model with multiple means of

[2] See the appendix for documentation.

payments.[3] The interest foregone by holding currency is an opportunity cost to private agents, and they are willing to incur real resource costs to avoid it. The resource cost of a money substitute is a social cost, while the interest cost associated with currency is not. Thus the private incentive to provide a substitute for government currency is greater than the social benefit, a point stressed by Wallace (1986). He has argued that a positive nominal interest rate provides a similar incentive to issue private banknotes (Wallace 1983). While banknotes employ virtually the same technology as government currency, stored value employs a very different technology but serves the same role—both are private substitutes for government currency.

The best policy in the model is one in which the nominal interest rate is zero—the Friedman rule for the optimum quantity of money (Friedman 1969). For a given positive nominal interest rate, however, Schreft (1992a) has shown that quantitative restrictions on the use of credit as a means of payment can improve welfare. The same is true for stored value as well, since stored value is just another form of credit as a means of payment; if nominal interest rates are positive, then the right kind of quantitative constraints on stored value cards, if practical, can improve welfare by preventing the most wasteful uses. A non-interest-earning reserve requirement on stored value liabilities is inferior to quantitative constraints because it imposes an inframarginal tax on users of stored value.

No attention is paid here to consumer protection or to the safety and soundness of bank stored value activities (see Blinder [1995] and Laster and Wenninger [1995]). The analysis presumes that stored value systems provide relatively fraud- and counterfeit-proof instruments. Historical instances of private issue of small-denomination bearer liabilities, such as the early nineteenth century U.S. “free banking” era, have raised concerns about fraudulent note issue (see Friedman [1960]). Williamson (1992) shows that a government monopoly in issuing circulating media of exchange may be preferable to laissez-faire private note issue due to asymmetric information about bank portfolios. Others argue that a system of private banknote issue can function rather well: see Rolnick and Weber (1983, 1984). In any case, current communications technologies and regulatory and legal restraints are quite different from those of the early nineteenth century. Whether the concerns of Friedman and Williamson are relevant to stored value cards is beyond the scope of this article; the focus here is the implication of the seigniorage tax for the private incentives to provide currency substitutes.

[3] For models of multiple means of payment see Prescott (1987), Schreft (1992a, 1992b), Marquis and Reffett (1992, 1994), Ireland (1994), Dotsey and Ireland (1995), English (1996), and Lacker and Schreft (1996).

1. STORED VALUE CARDS

Stored value cards—sometimes called "smart cards"—contain an embedded microprocessor and function much like currency for a consumer. Value is loaded onto a card at a bank branch, an automated teller machine (ATM), or at home through a telephone or computer hookup with a bank. Customers pay for the value loaded onto the card either by withdrawing funds from a deposit account or by inserting currency into a machine. Customers spend value by sliding it through a merchant's card reader, which reduces the card's balance by the amount of the purchase and adds it to the balance on the merchant's machine. Merchants redeem value at the end of the day through a clearing arrangement similar to those used for "off-line" credit card or ATM transactions. The merchant dials up the network and sends in the stored value, which is then credited to the merchant's bank account. More elaborate systems allow consumers to transfer value from card to card.

The value on a stored value card is a privately issued bearer liability. It is different from a check because the merchant does not bear the risk of insufficient funds in the buyer's account. It is different from a debit card in that the consumer hands over funds upon obtaining the stored value, while a debit card leaves the funds in the consumer's account until the transaction is cleared. Thus a debit card is a device for authorizing deposit account transfers, while a stored value card records past transfers.

For consumers, stored value cards can be more convenient than currency in many settings; some consumers are likely to find cards physically easier to handle than coins and paper notes. The technology could conceivably allow consumers to load value onto their cards using a device attached to their home computer, saving the classic "shoe leather" cost of bank transactions. For merchants, stored value cards offer many of the advantages of credit card sales. The merchant saves the trouble of handling coins and notes (banks often charge fees for merchant withdrawals and deposits of currency) and avoids the risk of employee theft. Stored value improves on the mechanisms used for credit and debit cards, however, because the merchant's device verifies the validity of the card, without costly and time-consuming on-line authorization. Thus stored value cards extend the electronic payments technology of credit card transactions to more time-sensitive settings where on-line authorization is prohibitive.

2. A MODEL OF CURRENCY AND STORED VALUE

This section describes a model in which stored value and currency both circulate in equilibrium. Monetary assets are useful in this model because agents are spatially separated and communication is costly. In the absence of stored value, agents use currency whenever shopping away from home, and the structure of

the model is reduced to a simple cash-in-advance framework. Stored value is a costly private substitute for currency, similar to costly trade credit in Schreft's (1992a, 1992b) models.

The model is a deterministic, discrete-time, infinite-horizon environment with a large number of locations and goods but no capital. At each location there are a large number of identical households endowed with time and a technology for producing a location-specific good. Each period has two stages. The first takes place before production and, in principle, allows any agent to trade any contingent claim with any other agent. During the second stage, production and exchange take place. One member of the household—the "shopper"—travels to other locations to acquire consumption goods. Simultaneously, the other member of the household—the "merchant"—produces location-specific goods and sells them to shoppers from other locations. Shoppers are unable to bring with them goods produced in their own location, so direct barter is infeasible.

The fundamental friction in this environment is that it is prohibitively costly for agents from two different locations to verify each other's identities. As a result, intertemporal exchange between agents from different locations is impossible, or, more precisely, not incentive compatible. Meetings between shoppers and merchants away from home thus effectively resemble the anonymous meetings of the Kiyotaki-Wright (1989) model. This provides a role for valued fiat currency. Because agents from the same location are known to each other, households can exchange arbitrary contingent claims with other households at the same location during the "securities market" in the first stage of each period. Households could travel to other locations during the securities market, but anonymity prevents meaningful exchange of intertemporal claims.

The stored value technology is a costly way of overcoming this friction. There are a large number of agents that are verifiably known to all—call them "issuers." They are price-takers and thus earn a competitive rate of return. Like all other agents they can travel to any other location during the securities market, but since their identities are known, they are able to issue enforceable claims. The claim people want to buy is one they can use in exchange in the goods market. The difficulty facing such an arrangement, however, is authenticating the claim to the merchant—in other words, the difficulty of providing the shopper with a means of communicating the earlier surrender of value. Issuers possess the technology for creating message-storage devices—"stored value cards"—that shoppers can carry and machines that can read and write messages on these devices. Issuers offer to install machines with merchants. These machines can read, verify, and write messages on shoppers' cards and can record and store messages.

In principle the stored value technology described here could be configured to communicate any arbitrary messages. In this setting, however, a very simple message space will suffice. The shopper's device carries a measure of

monetary value, and the merchant's device deducts the purchase price from the number on the shopper's card and adds it to the measure of value stored on the merchant's device. During the next day's securities market, the issuer visits the merchant and verifies the amount stored on the reader. The issuer sells stored value to households in securities markets and then redeems stored value from merchants during the next day's securities markets. Messages in this case function much like the tokens in Townsend's (1986, 1987, 1989) models of limited communication. From this perspective, currency and stored value can be seen as alternative communications mechanisms.

I will adopt a very simple assumption concerning the cost of the stored value technology. I will assume that the card-reader devices that merchants use are costless to produce but require maintenance each period and that only issuers have the expertise to perform this maintenance. The amount of maintenance required depends on the location in which the device is installed and is proportional to the real value of the transactions that were recorded on the device. Some locations are more suited to stored value systems than others. This assumption will allow stored value and currency to coexist in equilibrium, with currency used at the locations that are less well-suited for stored value. The proportionality of costs to value processed might reflect security measures or losses due to fraud that rise in proportion to the value of transactions. I make no effort to model such phenomena explicitly but merely take the posited cost function as given. There are no other direct resource costs. In particular, stored value cards themselves are costless.[4]

The stored value cost function here is quite simple and in many respects somewhat unrealistic. Merchants' devices and the communications networks used to "clear" stored value are, arguably, capital goods and should be represented as investment expenditures rather than input costs. I am abstracting from capital inputs here, but this seems appropriate in a model with no capital goods to begin with.[5] Another feature of my cost function is that cost is proportional to the value of the transaction. In practice, the resource cost of electronic storage and transmission might not vary much with the numerical value of the message: transmitting "ten" should not be much cheaper than transmitting "ten thousand." Thus it seems plausible that a communications system, once built, would be

[4] Indirect evidence suggests that the resource costs of stored value systems could be substantial. A variety of sources indicate that bank operating expenses associated with credit cards amount to around 3 or 4 percent of the value of credit card charge volume. This does not include the direct expenses of merchants such as the costs of card readers. Stored value systems may avoid some expenses such as the costs associated with credit card billing and the cost of on-line communications. On the other hand, the cost of stored value cards themselves are greater than the cost of traditional "magnetic strip" cards.

[5] See Ireland (1994), Marquis and Reffett (1992, 1994), and English (1996) for models with private payment technologies requiring capital goods.

equally capable of carrying large and small value messages.[6] The cost function I adopt is the simplest one that is sufficient to demonstrate the claim that the introduction of stored value cards can reduce economic welfare. One critical feature is that the relative opportunity costs of stored value and currency vary across transactions so that both potentially circulate in equilibrium. A second critical feature is that there are constant returns to scale in providing stored value at any given location so that competition among providers is feasible. It should become clear as I proceed that the results are likely to carry over to settings with more elaborate cost functions.

The assumed technology also implies that the choice between currency and stored value takes a particularly simple form. The cost of using currency is the interest foregone while it is in use. The cost of stored value is simply the resource cost described above. By assuming that government currency is costless, I am abstracting from many of the factors mentioned in the previous section such as physical convenience, currency handling costs, and employee theft. Costless currency simplifies the presentation without loss of generality. In the appendix I describe a model in which there are private costs of handling currency and show that Proposition 1 below still holds. One could also modify the model to include government currency costs, as in Lacker (1993). The appendix also contains a model in which stored value substitutes for other more costly means of payment such as checks or credit cards; Proposition 1 holds in that model as well.

I can now begin describing the model more formally.[7]

Households

Time is indexed by $t \geq 0$, and locations are indexed by z and h, where $z, h \in [0,1)$. For a typical household at location $h \in [0,1)$, consumption of good z at time t is given by $c_t(h,z)$, and labor effort is given by $n_t(h)$. Households are endowed with one unit of time that can be devoted to labor or leisure. The production technology requires one unit of labor to produce one unit of consumption good. Household preferences are

$$\sum_{t=0}^{\infty} \beta^t u(c_t(h), 1 - n_t(h)), \qquad c_t(h) = \inf_z c_t(h,z), \tag{1}$$

[6] Ireland (1994) studies a similar model in which cost is independent of the value of the transaction. Also see Prescott (1987) and English (1996). For a partial equilibrium model see Whitesell (1992).

[7] The model is most closely patterned after the environment in Schreft (1992a) and Lacker and Schreft (1996). The main difference is that here the alternative payments medium is used at a subset of locations by all shoppers visiting that location, rather than at all locations but by only a subset of shoppers visiting that location. Also, I allow a general cost function while Schreft's cost function is, for convenience, linear in distance.

where u is strictly concave and twice differentiable. Household preferences are thus Leontieff across goods. This assumption implies that the composition of consumption is unaffected by the relative transaction costs at different locations, which considerably simplifies matters. In addition, it implies that transaction costs at a given location are passed on entirely to shoppers, since demand at a given location is inelastic.[8] Since all goods will bear a positive price in equilibrium, we can assume without loss of generality that $c_t(h,z) = c_t(h)$ for all z.

In the securities market households acquire both currency and stored value. Since the units in which value is stored are arbitrary, there is no loss in generality in assuming that stored value is measured in units of currency. Thus one dollar buys one unit of stored value. Let $c_t^m(h,z)$ and $c_t^s(h,z)$ denote the consumption of good z at time t purchased with currency and stored value, respectively. Let $m_t(h)$ and $s_t(h)$ be the amount of currency and stored value acquired in the securities market at time t. Then the trading friction implies that

$$m_t(h) \geq \int_0^1 p_t c_t^m(h,z)dz \quad \text{and} \tag{2}$$

$$s_t(h) \geq \int_0^1 p_t^s(z) c_t^s(h,z)dz, \tag{3}$$

where p_t is the price of goods for currency and $p_t^s(z)$ is the price of goods for stored value at location z.[9]

Household h sells $y_t^m(h)$ units of output for currency and $y_t^s(h)$ units of output for stored value. In addition, they sell $y_t^i(h)$ units of output to issuers. Since issuers are well known, merchants are willing to sell to them on credit and accept payment in next period's securities market. Feasibility requires

$$y_t^m(h) + y_t^s(h) + y_t^i(h) \leq n_t(h). \tag{4}$$

At the end of the period, the household has $p_t^s(h)y_t^s(h)$ units of stored value on their card reader to be redeemed at $t+1$. Issuers pay interest at the nominal rate i_t, the market rate on one-period bonds, but deduct a proportional charge at rate $r_t(h)$ from the proceeds to cover their costs. Thus the household receives

[8] Allowing substitution between goods would imply that the composition of consumption varies with changes in relative transaction costs. Relative prices net of transaction costs would then vary across locations, destroying the symmetry in households' consumption and leisure choices. See Ireland (1994) and Dotsey and Ireland (1995) for models which relax the Leontieff assumption.

[9] In principle the price of goods for currency could vary across locations as well, but symmetry will ensure equality across locations. This is confirmed below: see (12). Note that shoppers do not receive explicit interest on stored value. Note also that I allow merchants to charge a different price for different payments instruments.

$[1 - r_t(h)](1 + i_t)p_t^s(h)y_t^s(h)$ units of currency at $t + 1$ for the stored value they have accepted.[10]

Households bring the following to the securities market: currency from the previous period's sales, stored value to be redeemed, maturing bonds, and any currency that might be left over from shopping in the previous period. Letting b_t be bond purchases at t, and τ_t be lump-sum taxes at t, households face the following budget constraint at time $t + 1$:

$$m_{t+1}(h) + s_{t+1}(h) + b_{t+1} + \tau_{t+1} \leq m_t(h) - \int_0^1 p_t c_t^m(h, z)dz + s_t(h) - \int_0^1 p_t^s(z)c_t^s(h, z)dz + (1 + i_t)b_t + p_t[y_t^m(h) + y_t^i(h)] + [1 - r_t(h)](1 + i_t)p_t^s(h)y_t^s(h). \quad (5)$$

Households maximize (1) subject to (2) through (5) and the relevant nonnegativity constraints, taking prices and interest rates as given.

Issuers

There are a large number of issuers at location zero, distinct from the households described above. Their preferences depend on their consumption $[c_t^i(z)]$ and leisure $(1 - n_t^i)$ according to

$$\sum_{t=0}^{\infty} \beta^t (c_t^i - n_t^i), \qquad c_t^i = \inf_z c_t^i(z). \quad (6)$$

Issuers sell $s_t(z)$ units of stored value per capita in securities market z at time t in exchange for $s_t(z)$ units of currency and redeem $s_t'(z)$ units of stored value per capita from merchants at market z at time $t + 1$ in exchange for $[1 - r_t(z)](1 + i_t)s_t'(z)$ units of currency. All of the stored value they issue will be spent by households and then redeemed from card readers in equilibrium, so issuers face the constraint

$$\int_0^1 s_t(z)dz = \int_0^1 s_t'(z)dz. \quad (7)$$

Maintenance of the devices on which the stored value at location z is redeemed requires $\gamma(z)s_t'(z)/p_t^s(z)$ units of labor effort, where $\gamma(z)$ is a continuous, strictly increasing function with $\gamma(0) = 0$. Total maintenance effort is therefore

$$n_t^i = \int_0^1 \gamma(z)s_t'(z)/p_t^s(z)dz. \quad (8)$$

[10] Note that the payment could be thought of as redemption at par plus a premium $[1 - r_t(h)](1 + i_t) - 1$. The form of merchants' payments to issuers depends on issuers' cost functions. The payment is proportional to the value redeemed because the issuers' costs are proportional to value redeemed. If costs were independent of value redeemed, the equilibrium payment would also be independent of value redeemed.

The alternative for issuers who are not active is to consume nothing [$c_t^i(z) = n_t^i = 0$], which can be interpreted as the proceeds of some alternative autarchic activity.

The family of an issuer consists of a worker and a shopper. The worker manages the stored value business, while the shopper travels around during the goods market period purchasing consumption. Since issuers are known to all, the shopper can buy on credit, paying $p_t c_t^i(z)$ in the location-z securities market at $t+1$ for goods purchased there at t. Excess funds are invested in bond holdings b_t^i. An issuer's securities market budget constraint is

$$p_t c_t^i + b_{t+1}^i \leq \int_0^1 s_{t+1}(z)dz + (1+i_t)b_t^i - \int_0^1 (1+i_t)[1-r_t(z)]s_t'(z)dz. \qquad (9)$$

Active issuers maximize (6), subject to (7) through (9) and the relevant nonnegativity constraints, by choosing consumption, bonds, labor effort, and the amount of stored value to issue and redeem at each location. Because there are a large number of issuers, competition between them will drive the utility of active issuers down to the reservation utility associated with inactivity. Issuers initially have no assets.

Government

The government issues fiat money M_t and one-period bonds B_t, collects lump-sum taxes T_t, and satisfies

$$M_{t+1} + B_{t+1} = M_t + (1+i_t)B_t - T_{t+1} \qquad (10)$$

for all t. The government sets a constant money growth rate $\pi = M_{t+1}/M_t - 1$, where $\pi \geq \beta - 1$.

Equilibrium

A symmetric monetary equilibrium consists of sequences of prices, quantities and initial conditions M_{-1} and $(1+i_{-1})B_{-1}$, such that households and issuers optimize, the lifetime utility of active issuers is equal to the lifetime utility of inactive issuers starting at any date, the government budget constraint (10) holds, and market clearing conditions hold for $M_t, B_t, y_t^m(h), y_t^s(h)$, and $y_t^i(h)$ for all t and z. I restrict attention to stationary equilibria, in which real magnitudes are constant over time. Where possible, time subscripts will be dropped from variables that are constant over time; variables refer to date t quantities unless otherwise noted.

The first-order necessary conditions for the issuer's maximization problem imply

$$r(z)(1+i) = \gamma(z)p_t/p_t^s(z). \qquad (11)$$

The left side of (11) is the nominal net return from issuing one dollar's worth of stored value at t to be redeemed at location z at $t+1$, with the proceeds invested in a bond maturing at $t+1$. Interest on the bond is paid over to the merchant, with a portion $r(z)$ of the payment deducted as a fee. The right side of (11) is the nominal cost of enough consumption goods to compensate the issuer for the disutility of maintaining the stored value device at location z. Thus condition (11) states that for stored value issuers marginal net revenue equals marginal cost at each location.

Merchants consider whether to sell output for currency or for stored value. The first-order necessary conditions for the household's maximization problem imply that if merchants are indifferent between accepting currency and stored value, then

$$p_t = [1 - r(z)](1 + i)p_t^s(z). \tag{12}$$

As a result, the last two terms in the household's budget constraint (5) simplify to $p_t n(h)$. Households at all locations face identical terms of trade between consumption and leisure despite the difference in transaction costs across locations. Consumption and labor supply are therefore identical across locations and the notation for h can be suppressed.

When shoppers consider whether to use currency or stored value to purchase consumption at location z, they compare the unit cost of the former, p_t, to the unit cost of the latter, $p_t^s(z)$. Using (11) and (12), shoppers use stored value if

$$p_t > p_t^s(z) = p_t[1 + \gamma(z)]/(1 + i). \tag{13}$$

Thus stored value is used where $\gamma(z) < i$. Because γ is strictly increasing, the boundary between the stored value and the currency locations can be written as a function $\zeta(i) \equiv \gamma^{-1}(i)$. Stored value will coexist with currency as long as i is less than $\gamma(1)$, the cost of stored value at the highest cost location. If $i \geq \gamma(1)$, then stored value drives out currency; in this case $\zeta(i)$ is one.

The total resource cost of stored value for a given nominal rate is

$$\int_0^{\zeta(i)} \gamma(z)dz \equiv \Gamma(i). \tag{14}$$

Steady-state equilibrium values of c and n can be found as the solutions to the first-order condition

$$u_1(c, 1 - n)/u_2(c, 1 - n) = 1 + [1 - \zeta(i)]i + \Gamma(i), \tag{15}$$

along with the feasibility condition

$$n = c[1 + \Gamma(i)]. \tag{16}$$

For a given nonnegative nominal interest rate, consumption and employment are determined by (15) and (16).[11] Assuming that neither leisure nor consumption are inferior goods, then $v(c, 1-n) \equiv u_1(c, 1-n)/u_2(c, 1-n)$ is decreasing in c and increasing in $1-n$. If in addition we assume that $v(c, 1-n)$ goes to infinity as c goes to zero and zero as $1-n$ goes to zero, then we are guaranteed an interior solution; the proof appears in the appendix. The real interest rate is $\beta^{-1} - 1$ in all equilibria, and the inflation rate is $\beta(1+i) - 1$.

Without stored value cards the economy has the same basic structure as a standard cash-in-advance model (Lucas and Stokey 1983) and can be obtained as a special case in which $\zeta(i)$ (and thus $\Gamma(i)$) equals zero.

3. THE WELFARE ECONOMICS OF STORED VALUE

An optimal steady-state allocation is defined by the property that no other feasible steady-state allocation makes at least one type of household better off without making some other type of household worse off. Two features of the model make optimality relatively easy to assess. Even though at some locations goods are sold for currency and at other locations goods are sold for stored value, households at all locations face identical terms of trade between consumption and leisure. As a result, all households at all locations will have the same lifetime utility in any given equilibrium. We can therefore focus our attention on the well-being of a representative household at a representative location. Second, because the lifetime utility of any given issuer is zero in all equilibria, we can effectively ignore the welfare of issuers when comparing equilibrium allocations. This just reflects the fact that issuers receive a competitive rate of return and are indifferent as to how they obtain it; constant returns to scale in providing stored value at any location implies that issuers earn no rents. Given these two features, we can compare alternative allocations by considering their effect on the lifetime utility of a representative household.

For the version of this economy without stored value, the welfare economics are well known. Households equate the marginal rate of substitution between consumption and leisure to $1+i$ rather than 1, the marginal rate of transformation, because consumption is provided for out of currency accumulated by working in the previous period, and currency holdings are implicitly taxed at rate i. Optimality requires that the marginal rate of substitution equal the marginal rate of transformation, which only holds if the nominal interest rate is zero. A positive nominal interest rate distorts household decisions,

[11] The reduced form structure in (15) and (16) is identical to Schreft (1992a) and Lacker and Schreft (1996), although the models are somewhat different. In Schreft's model households use credit when close to their own home location and currency when farther away; thus at every location shoppers from nearby use credit and shoppers from a distance use cash. In contrast, at some locations only stored value is used and at other locations only currency is used in my model. Also, in Schreft's model credit costs are linear in distance.

inducing substitution away from monetary activity (consumption) toward nonmonetary activity (leisure). The resulting welfare reduction is the deadweight loss from inflation in this model. Intuitively, inflation reduces the rate of return on currency, which causes consumers to economize on the use of currency. In a cash-in-advance economy they can do this only by consuming less of the goods whose purchase requires currency. The optimal monetary policy is to deflate at the rate of time preference, $\pi = \beta - 1$, so that the nominal interest rate is zero and the distortion in (15) is completely eliminated (Friedman 1969). Note that in the absence of stored value, inflation has no effect on the feasibility frontier (16).

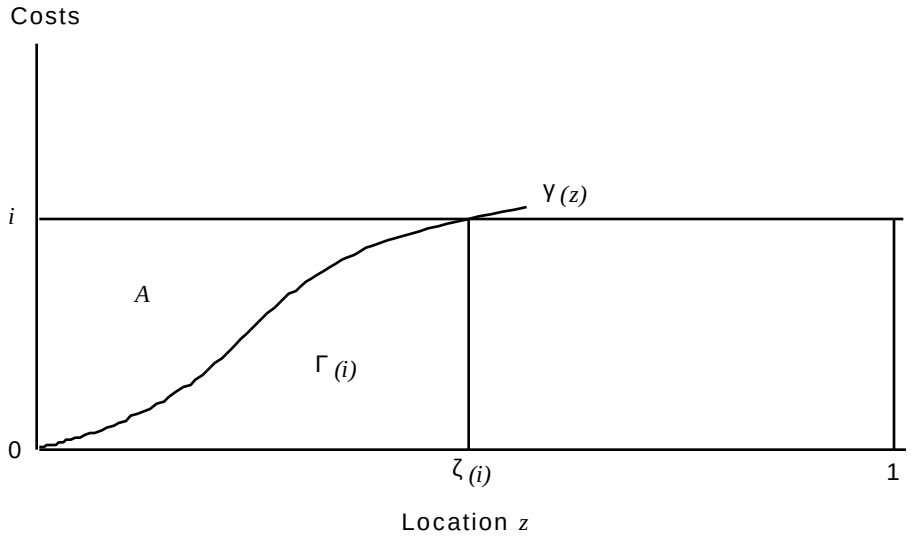
I will compare two steady-state equilibria with identical inflation rates, one with and one without stored value. Since the real rate is the same in all equilibria, the nominal rate is constant across equilibria as well. Stored value has two effects on a typical household's utility. The first is to alter the marginal rate of substitution between monetary and nonmonetary activities. The transaction cost associated with purchases using stored value at a given location $z < \zeta(i)$ is $\gamma(z)$, which is less than i, the private opportunity cost associated with using currency: see Figure 1. Thus stored value reduces the average transaction cost associated with consumption goods. This can be seen from (15), noting that the average cost of stored value, $\Gamma(i)/\zeta(i)$, is less than i. Stored value reduces the right side of (15) by the amount $\zeta(i)i - \Gamma(i)$, shown as the region A in Figure 1. Therefore, stored value cards reduce the distortion caused by inflation. By itself, this increases welfare. Note that the lower the total cost of stored value $\Gamma(i)$, holding constant i and $\zeta(i)$, the larger the welfare gain from stored value.

The second effect is through the feasibility constraint. Stored value cards involve real resource costs. Maintenance of the technology requires issuers' labor time, and consumption at every location must be diverted to compensate issuers for their effort. Currency requires no direct resource costs in this model.[12] The introduction of stored value shifts the consumption-leisure feasibility frontier (16) inward, since resources must be diverted to cover the real costs of stored value. The area under $\gamma(z)$ from zero to $\zeta(i)$ in Figure 1 is equal to $\Gamma(i)$, the real resource costs devoted to stored value activities. In contrast, the opportunity cost associated with currency (the area under i) is merely a transfer payment. By itself, the resource cost of stored value reduces welfare; a virtually costless government money is replaced by a costly private money. Note that the larger the total cost of stored value (holding constant i and $\zeta(i)$), the larger the reduction in welfare.

The net effect of stored value cards on economic welfare is indeterminate and depends on the structure of stored value costs across locations. Since we

[12] See the appendix for a model with positive private costs of using currency and Lacker (1993) for a similar model with positive government currency costs.

Figure 1 Opportunity Costs for Currency and Stored Value across Locations



are comparing two equilibria with the same nominal interest rate, we know that the marginal location has a cost of i. But conditions (15) and (16) tell us that the effect depends on the shape of the cost function. The benefit of stored value in (15) varies positively with the area A in Figure 1. The detrimental effect of stored value in (16) varies positively with $\Gamma(i)$. Both effects depend solely on $\Gamma(i)$ for any given i and $\zeta(i)$. If costs are low across most locations but then rise sharply—for example, if γ is quite convex—then $\Gamma(i)$ will be relatively small. In this case the negative effect through the feasibility condition will be small, and the positive effect through the marginal rate of substitution, $\zeta(i)i - \Gamma(i)$, will be large. If instead costs are large at most locations—for example, if γ is quite concave—then $\Gamma(i)$ is close to $\zeta(i)i$, the resource costs will be large, and the gain from reducing the marginal rate of substitution will be small. Thus the greater the convexity of costs across locations, the more likely it is that stored value cards improve economic welfare. This intuition is formalized in the following proposition. (The proof appears in the appendix.)

Proposition 1: *Fix i. Compare an economy with no stored value to an arbitrary stored value economy with a given ratio of stored value to currency, $\zeta(i)/[1 - \zeta(i)]$. There is a cutoff Γ^* [which depends on i and $\zeta(i)$] such that if $\Gamma(i) > \Gamma^*$, then welfare is lower in the stored value economy, and if $\Gamma(i) < \Gamma^*$, then welfare is higher in the stored value economy.*

The principle described in Proposition 1 appears to be quite general. In any model in which there is a deadweight loss due to the inflation tax, a private substitute for currency will reduce the base on which the tax is levied. The cost of the substitute must be less than the tax it evades—otherwise it would not be introduced. With the tax rate (the nominal interest rate) held constant, the incidence of the tax is lower and so the deadweight loss associated with that distortion will fall. Thus in any model we should expect that private substitutes for fiat money help reduce the burden of a given inflation rate.

The negative welfare effect of stored value cards would seem to generalize as well. Absent market imperfection, participants will adopt stored value if their collective private benefits exceed their collective private costs. But their net benefits differ from social net benefits in two ways; the capture of seigniorage is not a social benefit, and they do not bear the governmental cost of currency provision. The gap between the nominal interest rate and the government's per-dollar cost of providing currency thus represents the *excess* incentive to implement currency substitutes.

More realistic or elaborate models would also display the principle described in Proposition 1. For example, if stored value costs do not depend directly on the value of messages, then the fee an issuer collects from a merchant would be independent of the merchant's sales.[13] The relevant cost comparison is then between the seigniorage tax, which is proportional to production, and the fixed fee, which is not. In this case the set of locations using stored value would vary with equilibrium consumption, instead of being independent of consumption as in the model above. Nevertheless, stored value would reduce transaction costs where it is used, and the benefit of a reduced inflation tax burden would have to be weighed against the added resource costs.

In the absence of stored value, inflation is costly in the model because it distorts the choice between monetary and nonmonetary activities. Some economists have suggested that an important cost of inflation is that it encourages costly private credit arrangements as substitutes for government money (see Ireland [1994], Dotsey and Ireland [1995], Lacker and Schreft [1996], and Aiyagari, Braun, and Eckstein [1995]). Stored value could reduce this cost of inflation by displacing even more costly means of payment such as checks or credit cards. The social benefit of stored value would then also include the reduction in payments cost for some transactions. This benefit would be larger, the smaller the cost of stored value. But again, the benefit of stored value would have to be weighed against the resource cost of substituting stored value for virtually costless currency. As long as stored value in part substitutes for currency, Proposition 1 again emerges; the lower the average cost of stored value cards, the greater the gain from displacing more costly means of payments, and

[13] The payments technologies in Ireland (1994) and English (1996) have this property.

the smaller the resources diverted to stored value systems. (A straightforward extension of the model demonstrating this result is described in the appendix.)

As I mentioned earlier, stored value seems likely to offer consumers and merchants greater physical convenience or some other advantages over currency that are not captured in the model presented above. Such advantages would not alter the basic feature of Proposition 1, however. If merchants find stored value less costly to handle than currency, their savings will presumably be reflected in their willingness to pay stored value issuers; the social benefit of stored value to merchants will be reflected in issuers' revenues. Similarly, if consumers find stored value more convenient than currency, they should be willing to pay, either directly or indirectly, and the benefit of stored value to consumers will be reflected in issuers' revenues. The greater convenience of stored value cards will provide issuers with added incentive to provide stored value, but the nominal interest rate (minus the government currency cost) would *still* constitute a source of private return to issuing stored value that is not matched by any social benefit of replacing currency. To demonstrate this point, the appendix describes a simple extension of the model in which there are private costs to handling currency and shows that Proposition 1 again holds true.

What if consumers earn interest on their stored value cards, a possibility that appears to be technologically feasible? Could this upset the conclusions of Proposition 1? In the model, merchants earn interest on stored value balances, although they pay some of this interest back to issuers in the form of redemption fees. Stored value yields no explicit interest for shoppers. An equivalent scheme would be for shoppers to earn interest on stored value but face higher prices at locations that accept stored value. The interest earnings would more than compensate shoppers for the higher price at those locations. Stored value would again be used at locations at which the resource cost does not exceed the opportunity cost of using currency.

What if stored value completely displaces currency? In this case the meaning of the nominal interest rate in the model becomes somewhat ambiguous; because currency does not circulate, it might not serve as the unit of account. Nonetheless, the difference between the real return on bonds and the real return on stored value liabilities will not exceed the marginal cost of providing stored value.[14] Unless this difference is less than the government's cost of providing currency, the principle underlying Proposition 1 still applies.

Bryant and Wallace (1984) have argued that different rates of return on different government liabilities can be justified as an optimal tax. If all sources

[14] This could happen in one of two ways. Currency could remain the unit of account—a "ghost" money. Issuers would pay interest to consumers on stored value, and issuers' net interest margin would not exceed the marginal cost of stored value. Alternatively, stored value could become the unit of account, in which case the nominal interest rate would fall to the marginal cost of stored value.

of government revenue require distortionary taxation, then it may be beneficial to raise some revenues from the taxation of currency holders. This consideration is not captured in the model described above. The seigniorage tax merely finances interest payments on government bonds. The reduction in seigniorage revenues was offset by a reduction in outstanding government debt or an increase in lump-sum taxes, keeping the nominal rate constant. If instead the loss of seigniorage had to be recovered by raising other distortionary taxes, it would strengthen the case against stored value. The additional deadweight burden of the compensating tax increases would have to be added to the resource cost of stored value. Similarly, nonzero government expenditures financed in part through seigniorage would not change the basic features of Proposition 1.

4. POLICY

In the presence of one distortionary tax a second distortion can sometimes improve economic welfare. A positive nominal interest rate is a distortionary tax on holders of government currency. In the laissez-faire regime with stored value, welfare can be lower because of the costs of stored value, which suggests that constraints on the issue of stored value cards might be welfare-enhancing. This turns out to be true. Restrictions on the issue of stored value can improve welfare, in this second-best situation, by reducing the costly displacement of government currency.[15]

Consider first a simple quantitative restriction on the use of stored value. Imagine a legal restriction that limits the quantity of stored value used by households, or, equivalently, that limits the locations at which stored value is accepted. Households can only make a fraction η of their purchases using stored value, where the government sets η between 0 and $\zeta(i)$. Households will continue to use stored value where it is most advantageous to do so—at locations 0 through η. At locations η through 1, households use currency. Equilibrium is still characterized as the solution to (15) and (16), except that $\zeta(i)$ is replaced by η. For a given nominal interest rate, consumption and employment are determined as the solutions to

$$u_1(c, 1-n)/u_2(c, 1-n) = 1 + (1-\eta)i + \int_0^\eta \gamma(z)dz \quad \text{and} \tag{17}$$

$$n = c[1 + \int_0^\eta \gamma(z)dz]. \tag{18}$$

Define $c(\eta)$ and $n(\eta)$ as the solutions to (17) and (18) for $\eta \in [0, \zeta(i)]$, and $v(\eta) \equiv u(c(\eta), 1 - n(\eta))$. The function $v(\eta)$ is the equilibrium utility of a

[15] This depends, of course, on finding a practical way to restrict the quantity of stored value issued. I do not address this issue here.

representative household under the constraint that no more than a fraction η of purchases can be made using stored value.

Proposition 2: *$v'[\zeta(i)] < 0$; therefore there is a binding restriction $\eta < \zeta(i)$ on stored value under which steady-state utility is higher than under the laissez-faire regime.*

Reducing η marginally below $\zeta(i)$ has two effects. The direct effect via the resource constraint (18) is to eliminate the most costly uses of stored value, allowing greater consumption at each level of employment. The increase in utility at $\eta = \zeta(i)$ is proportional to $\gamma(i)$. The second effect via the marginal rate of substitution is to substitute currency (transaction cost i) for stored value (transaction cost $\gamma(\eta)$). The fall in utility is proportional to $i - \gamma(\eta)$, which vanishes at $\eta = \zeta(i) = \gamma^{-1}(i)$, since at the margin stored value and currency bear the same transaction cost. The first-order resource savings dominates the negligible increase in the burden of inflation. The net effect of decreasing η is positive. Therefore there must be a value $\eta < \zeta(i)$ that results in higher steady-state utility than the laissez-faire regime. Note that Proposition 2 holds whether or not the stored value equilibrium is worse than the no-stored-value equilibrium; even if stored value is welfare-enhancing, quantitative constraints would still be worthwhile.[16]

An alternative method of restraining stored value is to impose a reserve requirement. Issuers are required to hold currency equal to a fraction δ of their outstanding stored value liabilities. Issuers earn $(1 - \delta)i$ rather than i on their assets, and the cost of foregone earnings, δi, is passed on to merchants and ultimately to consumers. Stored value is used at fewer locations for any given interest rate—only where $\gamma(i) < (1-\delta)i$. Raising the reserve requirement from zero reduces the amount of resources diverted to stored value. By itself this increases utility by easing the feasibility frontier in (18). The first-order condition (15) becomes

$$u_1(c, 1-n)/u_2(c, 1-n) = 1 + [1 - \zeta(i, \delta)]i + \int_0^{\zeta(i,\delta)} (1-\delta)^{-1}\gamma(z)dz, \qquad (19)$$

where $\zeta(i, \delta) \equiv \gamma^{-1}[(1-\delta)i]$. Increasing the reserve requirement now has two effects on the marginal rate of substitution. First, raising δ expands the set of

[16] Proposition 2 parallels Proposition 2 in Schreft (1992a). In Schreft's model the government issues no bonds and government expenditure depends on the seigniorage collected. Lowering the constraint holding the inflation rate constant increases the demand for money and thus government expenditures. Schreft's proposition requires the condition that government expenditures rise by less than the resource cost of payments services falls when the constraint is tightened. This condition is unnecessary if government spending is held constant and instead the bond supply or lump-sum taxes vary across equilibria. There are no government expenditures in my model. Note that Proposition 2 depends crucially on the continuity of stored value costs across locations. If there were a discrete jump in the function γ at $\zeta(i)$, then Proposition 2 might fail to hold.

locations at which currency is used instead of lower-cost stored value. This effect operates through $\zeta(i, \delta)$ in (19). Second, raising δ increases the transaction cost at all locations at which stored value is used. This effect increases the integrand in the last term in (19) and does not vanish at $\delta = 0$. The first effect is identical to the effect of decreasing the quantitative constraint η in (17). The quantitative constraint does not involve the second effect in (19), since it leaves inframarginal stored value users unaffected. In contrast, the reserve requirement imposes a tax on all stored value users. Therefore, a quantitative constraint is superior to a reserve requirement in this environment. A reserve requirement improves welfare only if the second effect in (19) does not dominate.[17]

5. CONCLUDING REMARKS

When the nominal interest rate is positive, there is an incentive to develop substitutes for government currency. This incentive is likely to exceed the social benefit of such substitutes because the private opportunity cost of holding currency is larger than the social cost of providing currency. Although stored value can lower the opportunity cost of payments media for inframarginal consumers, the real resources diverted to stored value are wasteful from society's point of view. For a given nominal interest rate, stored value cards are good or bad for welfare depending upon whether the average cost of stored value is below or above a certain cutoff. Quantitative restrictions on stored value can be socially beneficial, in this second-best situation, because they reduce the amount of resources absorbed by the most costly stored value applications. I do not claim to show that such restrictions can be easily implemented—only that if such restrictions were practical, they would enhance welfare.

Wallace (1983) pointed out that the U.S. government has effectively prohibited the private issue of paper small-denomination bearer notes such as bank notes. In the absence of such a ban, he argued, private intermediaries could issue perfect substitutes for government currency backed by default-free securities such as U.S. Treasury bills. In this case one of two things could occur. Either the nominal rate of return would not exceed the marginal cost of such intermediation, which he reckoned at close to zero, or government currency would cease to circulate. Stored value cards are just another way of issuing small-denomination bearer liabilities. As a corollary to Wallace's thesis, then, we should expect one of two things to happen. Either the nominal interest rate will not exceed the marginal cost of an additional dollar's worth of stored value, or government currency will cease to circulate. All I have added to Wallace's

[17] A reserve requirement equal to $1 - \gamma_g/i$ could be imposed (where γ_g is the government currency cost) as an experiment to determine whether the resource costs of stored value exceed the direct benefits, i.e., whether $\gamma(z) > \gamma_g$. This experiment would not answer the question posed by Proposition 1, however.

argument is the observation that since stored value employs a technology that is different from, and potentially far more costly than, the government currency it would replace, it is possible that either outcome could reduce economic welfare.

The restrictions that have prevented private paper substitutes for currency were in place since at least the end of the Civil War. As Wallace (1986) notes, "(i)f there is a rationale for that policy . . . then it would seem that it would apply to other payments instruments that potentially substitute for the monetary base." These longstanding restrictions on paper note issue evidently have been repealed.[18] Current policy thus avoids the inconsistency of allowing electronic substitutes for government currency while preventing paper substitutes.

APPENDIX

Government Currency Costs

The Annual Report of the Director of the Mint reports the coinage cost per $1,000 face value for every denomination of coin, along with the number manufactured. For 1993 (the latest year available) this yields a coin manufacturing cost of $166.2 million (31.3 percent of face value). For coin operating cost the 1994 PACS Expense Report lists total cost of coin service of $14.7 million (Board of Governors 1994b). Total government cost of coin is the sum of manufacturing and operating costs, or $180.9 million. U.S. Treasury Department Bulletin reports coin in circulation on December 31, 1993, as $20.804 billion. For coin, therefore, the cost per dollar outstanding is $0.008695.

For currency, governmental cost for 1993 is the sum of Federal Reserve Bank operating expenses of $123.7 million (Board of Governors 1994b), and the Reserve Bank assessment for U.S. Treasury currency expenses of $355.9 million (Board of Governors 1994a). Total cost for currency is thus $479.6 million, or $0.001394 per dollar outstanding, based on $343.925 billion in currency outstanding at the end of 1993 (Board of Governors 1994a).[19]

Combining currency and coin costs yields a total of $660.5 million for 1993. The total value of currency and coin in circulation was $36.729 billion. The total government cost of coin and currency per dollar outstanding is therefore $0.001798.

[18] Title VI of the Community Development Banking Act, P.L. 103-325 (1994), repealed all restrictions on note issue by national banks except the 1/2 percent semi-annual tax on outstanding notes. Section 1904(a) of the Tax Reform Act of 1976 repealed the 10 percent tax on note issue by corporations other than national banks. De facto restrictions by bank regulators may still prevent private note issue.

[19] For more on the government's cost of currency, see Lacker (1993).

Proofs

Existence: Define $c(y)$ and $n(y)$ as the joint solutions to $u_1(c, 1-n) = qu_2(c, 1-n)$ and $qc + (1-n) = y$. Here q is the marginal rate of substitution between consumption and leisure, the right-hand side of (15). The assumptions on preferences imply that $c(y)$ and $1 - n(y)$ are unique, strictly positive for $y > 0$, continuous, and monotone increasing. Assume $r > 0$, where r is the marginal rate of transformation between consumption and leisure, the bracketed term in (16). Then $rc(y) + [1 - n(y)]$ is strictly increasing in y, there exists a unique y such that $rc(y) + [1 - n(y)] = 1$, and thus $n(y) = rc(y)$.

Proposition 1: Define $c(q, r)$ and $n(q, r)$ as the unique solutions to

$$u_1(c, 1 - n) = qu_2(c, 1 - n)$$

$$n = rc,$$

where attention is restricted to $r \geq 1$ and $q \geq r$. Define $V(q, r) = u(c(q, r), 1 - n(q, r))$. It is easy to show that since neither leisure nor consumption is an inferior good, V is strictly decreasing in q and r.

The first-best allocation has a nominal interest rate of zero, so $q = r = 1$. Economies with positive nominal rates but no stored value have $q = 1 + i > 1$ and $r = 1$. For given i and $\zeta(i), q$ and r vary positively with the aggregate $\Gamma(i)$. This amounts to varying average cost holding marginal cost constant at location $z = \zeta(i)$. Note that $\Gamma(i)$ can lie anywhere in the interval $(0, \zeta(i)i)$. Define $w(\Gamma) = V(1+(1-\zeta(i))i+\Gamma, 1+\Gamma)$. Then $w(0) = V(1+(1-\zeta)i, 1) > V(1+i, 1)$, and $w(\zeta(i)i) = V(1 + i, 1 + \zeta(i)i) < V(1 + i, 1)$. Since $w(\Gamma)$ is continuous and strictly decreasing in Γ, it follows immediately that there exists a unique Γ^* for which $w(\Gamma^*) = V(1 + i, 1)$.

Proposition 2: Define $q(\eta)$ as the right side of (17), and $r(\eta)$ as the bracketed term in (18). With $v(\eta) \equiv V(q(\eta), r(\eta))$, we have $\lim_{\eta \to \zeta(i)} v'(\eta) = \lim_{\eta \to \zeta(i)} [V_q q' + V_r r'] = V_q[\gamma(\zeta(i)) - i] + V_r \gamma(\zeta(i)) = V_r \gamma(\zeta(i)) < 0$, since $\gamma(\zeta(i)) = i$ and $V_r < 0$.

A Model in Which Stored Value Substitutes for Other Means of Payment

In this section I describe a simple extension of the model in which there are three means of payment: government currency, stored value, and another costly private means of payment. The latter can be thought of as checks or credit cards and is supplied by an industry with the same general properties as the stored value sector. Locations will now be indexed by z_1 and z_2. Stored value costs depend only on z_1 according to $\gamma(z_1)$. The cost of checks depends only on z_2 according to a continuous and monotone increasing function $\chi(z_2)$. To simplify the exposition, I will abstract from the effect of inflation on labor supply and assume that labor is supplied inelastically: $n \equiv 1$ and $u(c, 1 - n) \equiv u(c)$.

Inflation is inefficient in this version of the model solely because it diverts resources to the production of money substitutes. Consumption equals output (which is fixed and equal to 1) minus the resource costs of checks and stored value. In the absence of stored value, shoppers use currency at locations where $\chi(z_2) > i$, and use checks where $\chi(z_2) < i$. Without stored value, then, the cost of inflation is

$$\int_0^1 \int_0^{\chi_{-1}(i)} \chi(z_2) dz_2 dz_1.$$

Shoppers will use stored value at locations where $\gamma(z_1) < \min[\chi(z_2), i]$. With stored value, the consumption diverted to alternative monies is

$$\int_{S(i)} \gamma(z_1) dz_2 dz_1 + \int_{C(i)} \chi(z_2) dz_2 dz_1,$$

$$\text{where } S(i) \equiv \{(z_1, z_2) \text{ s.t. } \gamma(z_1) < \text{ MIN}[i, \chi(z_2)]\}$$

$$\text{and } C(i) \equiv \{(z_1, z_2) \text{ s.t. } \chi(z_2) < \text{ MIN}[i, \gamma(z_1)]\}.$$

As in the model in the text, stored value wastes resources: at locations described by $z_1 < \gamma^{-1}(i)$ and $z_2 > \chi^{-1}(i)$, stored value costs are incurred where formerly (costless) currency was in use. At these locations, the greater the average cost of stored value the greater the resource diversion. However, at locations described by $z_2 < \chi^{-1}(i)$ and $\gamma(z_1) < \chi(z_2)$, stored value substitutes for more costly check use. In this region the resource savings is larger, the smaller the average cost of stored value. Whether the resource savings from displacing checks outweighs the resource cost of displacing currency depends on the stored value cost function. It is straightforward to show that there is once again a cutoff value; if average stored value costs are below the cutoff, stored value is welfare-enhancing, while if average cost is greater than the cutoff, stored value reduces welfare.

A Model with Costs of Handling Currency

This section describes an extension of the model in which handling currency is costly to merchants and shows that Proposition 1 also holds in this extended economy. (The same is true of an extended model in which handling currency is costly to shoppers, but that model is omitted here.) Suppose then that accepting currency as payment requires time-consuming effort on the part of merchants. For convenience, I will assume that the time requirement is proportional to the real value of currency handled and is equal to

$$\frac{\alpha p_t y_t^m(h)}{p_t},$$

where the parameter $\alpha > 0$. The feasibility condition (4) now becomes

$$y_t^m(h) + \alpha y_t^m(h) + y_t^s(h) + y_t^i(h) \leq n_t(h). \qquad (4')$$

One unit of labor devoted to goods sold for currency now yields $p_t/(1+\alpha)$ units of currency at the beginning of period $t+1$. One unit of labor devoted to goods sold to an issuer must provide the same yield, so in (5) $y_t^i(h)$ is replaced with $y_t^i(h)/(1+\alpha)$. In the issuer's budget constraint, $p_t c_t^i(z)$ is replaced by $p_t c_t^i(z)/(1+\alpha)$. The first-order condition from the issuer's problem becomes

$$r(z)(1+i) = \frac{\gamma(z)p_t}{(1+\alpha)p_t^s(z)}. \tag{11$'$}$$

One unit of labor devoted to cash sales yields $p_t/(1+\alpha)$ units of currency at $t+1$. Therefore (12) becomes

$$p_t = (1+\alpha)[1-r(z)](1+i)p_t^s(z). \tag{12$'$}$$

The last two terms in the household's budget constraint (5) now simplify to $p_t n_t(h)/(1+\alpha)$.

Shoppers now use stored value if and only if $(1+\alpha)(1+i) > [1+\gamma(z)]$. Define $\zeta(i)$ by $(1+\alpha)(1+i) = (1+\gamma(\zeta(i)))$. At this point I make a minor modification to the model. The preferences of issuers are altered so that goods from different locations are perfect substitutes:

$$c_t^i = \int_0^1 c_t^i(z)dz.$$

With this modification, the feasibility condition for this model simplifies to

$$n = c\{(1+\alpha)[1-\zeta(i)] + \Gamma(i)\}, \tag{16$'$}$$

reflecting the fact that both currency handling costs and stored value costs affect the aggregate resource constraint. The first-order condition for this model is

$$\frac{u_1(c,1-n)}{u_2(c,1-n)} = (1+\alpha)\{1+[1-\zeta(i)]i+\Gamma(i)\} \equiv q. \tag{15$'$}$$

Once again the optimal monetary policy is the Friedman rule, but now stored value circulates even when the nominal interest rate is zero, since in some applications stored value is less costly than currency [$\gamma(z) < \alpha$]. It is easy to demonstrate that Proposition 1 holds in this economy as well: for any fixed positive nominal interest rate there exists a cutoff Γ^* such that the stored value economy Pareto-dominates the economy without stored value if and only if $\Gamma(i) < \Gamma^*$.

REFERENCES

Aiyagari, S. Rao, Toni Braun, and Zvi Eckstein. "Transactions Services, Inflation, and Welfare." Manuscript. 1995.

Blinder, Alan. Statement before the Subcommittee on Domestic and International Monetary Policy of the Committee on Banking and Financial Services, U.S. House of Representatives, October 11, 1995.

Board of Governors of the Federal Reserve System. *80th Annual Report,* 1993. Washington: Board of Governors, 1994a.

_____. *PACS Expense Report,* 1993. Washington: Board of Governors, 1994b.

Bryant, John, and Neil Wallace. "A Price Discrimination Analysis of Monetary Policy," *Review of Economic Studies,* vol. 51 (April 1984), pp. 279–88.

Dotsey, Michael, and Peter Ireland. "The Welfare Cost of Inflation in General Equilibrium." Manuscript. 1995.

English, William B. "Inflation and Financial Sector Size," Federal Reserve Board Finance and Economics Discussion Series No. 96–16 (April 1996).

Friedman, Milton. "The Optimum Quantity of Money," in *The Optimum Quantity of Money and Other Essays*. Chicago: Aldine, 1969.

_____. *A Program for Monetary Stability*. New York: Fordham Univ. Press, 1960.

Ireland, Peter N. "Money and Growth: An Alternative Approach," *American Economic Review,* vol. 84 (March 1994), pp. 47–65.

Kiyotaki, Nobuhiro, and Randall Wright. "On Money as a Medium of Exchange," *Journal of Political Economy,* vol. 97 (August 1989), pp. 927–54.

Lacker, Jeffrey M. "Should We Subsidize the Use of Currency?" Federal Reserve Bank of Richmond *Economic Quarterly,* vol. 79 (Winter 1993), pp. 47–73.

_____, and Stacey L. Schreft. "Money and Credit as Means of Payment," *Journal of Monetary Economics,* vol. 38 (August 1996), pp. 3-23.

Laster, David, and John Wenninger. "Policy Issues Raised by Electronic Money." Paper presented to the Conference on Digital Cash and Electronic Money, held at the Columbia Business School, April 21, 1995.

Lucas, Robert E., Jr., and Nancy L. Stokey. "Optimal Fiscal and Monetary Policy in an Economy without Capital," *Journal of Monetary Economics,* vol. 12 (July 1983), pp. 55–93.

Marquis, Milton H., and Kevin L. Reffett. "New Technology Spillovers into the Payment System," *Economic Journal,* vol. 104 (September 1994), pp. 1123–38.

_____. "Capital in the Payments System," *Economica,* vol. 59 (August 1992), pp. 351–64.

Prescott, Edward C. "A Multiple Means of Payment Model," in William A. Barnett and Kenneth J. Singleton, eds., *New Approaches to Monetary Economics*. Cambridge: Cambridge Univ. Press, 1987.

Rolnick, Arthur J., and Warren E. Weber. "The Causes of Free Bank Failures: A Detailed Examination," *Journal of Monetary Economics*, vol. 14 (November 1984), pp. 267–92.

_____. "New Evidence on the Free Banking Era," *American Economic Review*, vol. 73 (December 1983), pp. 1080–91.

Schreft, Stacey L. "Welfare-Improving Credit Controls," *Journal of Monetary Economics*, vol. 30 (October 1992a), pp. 57–72.

_____. "Transaction Costs and the Use of Cash and Credit," *Economic Theory*, vol. 2 (1992b), pp. 283–96.

Townsend, Robert M. "Currency and Credit in a Private Information Economy," *Journal of Political Economy*, vol. 97 (December 1989), pp. 1323–44.

_____. "Economic Organization with Limited Communication," *American Economic Review*, vol. 77 (December 1987), pp. 954–71.

_____. "Financial Structures as Communications Systems," in Colin Lawrence and Robert P. Shay, eds., *Technological Innovation, Regulation, and the Monetary Economy*. Cambridge, Mass.: Ballinger, 1986.

Wallace, Neil. "The Impact of New Payment Technologies: A Macro View," in Colin Lawrence and Robert P. Shay, eds., *Technological Innovation, Regulation, and the Monetary Economy*. Cambridge, Mass.: Ballinger, 1986.

_____. "A Legal Restrictions Theory of the Demand for 'Money' and the Role of Monetary Policy," Federal Reserve Bank of Minneapolis *Quarterly Review*, vol. 7 (Winter 1983), pp. 1–7.

Whitesell, William C. "Deposit Banks and the Market for Payment Media," *Journal of Money, Credit, and Banking*, vol. 24 (November 1992), pp. 483–98.

Williamson, Stephen D. "Laissez-Faire Banking and Circulating Media of Exchange," *Journal of Financial Intermediation*, vol. 2 (June 1992), pp. 134–67.

Monetary Policy and Long-Term Interest Rates

Yash P. Mehra

The standard view of the transmission mechanism of monetary policy assigns a key role to long-term interest rates. According to this view, a monetary policy tightening pushes up both short and long interest rates, leading to less spending by interest-sensitive sectors of the economy and therefore to lower real growth. Conversely, a monetary easing results in lower interest rates that stimulate real growth. An open question in discussions of this view is whether monetary policy has significant empirical effects on long-term interest rates.[1]

In this article, I provide new evidence on the quantitative effect of monetary policy on the long-term interest rate. The federal funds rate is used as a measure of monetary policy.[2] The work extends the previous research in

The author thanks Peter Ireland, Roy Webb, and Marvin Goodfriend for their many helpful comments. The views expressed are those of the author and do not necessarily represent those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] In most previous studies using an interest-rate-based measure of monetary policy, a short-term money market rate (the three-month or one-year Treasury bill rate) is used. Most of those studies surveyed recently in Akhtar (1995) find significant and large effects of short rates on long rates. In those studies the long-run response of nominal long rates ranges from about 22 to 66 basis points for every one percentage point change in nominal short rates. However, there is considerable skepticism about the reliability and interpretation of those effects. One main reason for such skepticism is even though monetary policy has its strongest effect on a short-term money market rate, the latter is also influenced by nonmonetary forces. Hence changes in short rates do not necessarily reflect changes in the stance of monetary policy.

There are a few other empirical studies that use the federal funds rate as a measure of monetary policy. But most of those studies examine the effect of policy on the long rate in a bivariable framework. In such studies the estimated impact of policy on the long rate is quantitatively modest and temporally unstable (see Akhtar 1995, Table 3). An exception is the recent work in Mehra (1994) which uses a multivariable framework and finds a significant effect of the real funds rate on the long rate. However, Mehra (1994) does not investigate the robustness of those results to alternative specifications or to different sample periods.

[2] Recent research has shown that the federal funds rate is a good indicator of the stance of monetary policy (Bernanke and Blinder 1992; Bernanke and Mihov 1995).

two main directions. First, following Goodfriend's (1993) description of funds rate policy actions, it distinguishes empirically between the long- and short-run sources of interaction between the funds rate and the long rate; this distinction is absent in previous studies. Second, the analysis in Goodfriend also suggests that the near-term effects of funds rate policy actions on the long rate may be quite variable. Hence the work examines the temporal stability of such effects, an issue also virtually ignored in previous research.

The empirical work focuses on the behavior of the nominal yield on ten-year U.S. Treasury bonds during the period 1957Q1 to 1995Q2. The results here indicate that the bond rate moves positively with the funds rate in the long run. However, this comovement arises because the bond rate automatically moves with trend inflation (the Fisher relation) and the Federal Reserve (Fed) keeps the level of the funds rate in line with the going trend rate of inflation (the long-run Fed reaction function). Apart from the correlation that occurs through the inflation channel, I find that empirically there is no other source of long-run interaction between the bond rate and the funds rate. This result arises because the bond rate's other component—the expected long real rate—is mean stationary and therefore unrelated to the level of the funds rate in the long run. These results have the policy implication that monetary policy can permanently lower the bond rate only by lowering the trend rate of inflation.

The short-run stance of monetary policy is measured by the spread between the funds rate and the trend rate of inflation (the funds rate spread). The results indicate that movements in the funds rate spread have a statistically significant effect on the bond rate and that the magnitude of its near-term effect on the bond rate has increased significantly since 1979.[3] In the pre-1979 period, the bond rate rose anywhere from 14 to 29 basis points whenever the funds rate spread widened by one percentage point. In the post-1979 period, however, the estimate of its near-term response ranges from 26 to 50 basis points.

The short-run results thus suggest that, ceteris paribus, a monetary policy tightening measured by an increase in the funds rate spread does result in higher bond rates in the short run, in line with the traditional view of the transmission mechanism. However, this increase in the short-run sensitivity of the bond rate to policy actions may itself be due to the way the Fed has conducted its monetary policy since 1979. The Fed's post-1979 efforts to bring down the trend rate of inflation, coupled with lack of credibility, may have amplified the near-term effects of funds rate changes on the bond rate.

[3] In this article near-term effects refer to responses of the bond rate to recent past values of the funds rate spread. The immediate effect is the response to the one-period lagged value of the spread and the near-term effect is the cumulative response to all such past values. What I call the near-term effect is sometimes referred to as the long-run effect in previous studies. As indicated later, I use the long-run effect to measure the effect that arises from the existence of equilibrium or trending relationships among nonstationary variables.

The plan of this article is as follows. Section 1 presents the methodology that underlies the empirical work, Section 2 contains empirical results, and concluding observations are in Section 3.

1. THE MODEL AND THE METHOD

This section describes the methodology that underlies the empirical work in this article.

The Fisher Relation, the Bond Rate, and the Federal Funds Rate

In order to motivate the empirical work, I first describe how monetary policy may affect the bond rate in the short run and the long run.[4] As indicated before, the federal funds rate is used as a measure of the stance of monetary policy. Thus a monetary policy action is defined as a change in the funds rate, and a monetary policy strategy is defined as the reaction function that would lead to policy actions. The Fisher relation for interest rates provides a convenient framework within which effects of policy actions can be described. The Fisher relation is

$$BR_t = rr_t^e + \dot{p}_t^e, \tag{1}$$

where BR_t is the bond rate, rr_t^e is the expected long real rate, and $\dot{p}_t^e$ is the expected long-term inflation rate. Equation (1) relates the bond rate to expectations of inflation and the real rate.

The Fisher relation indicates that policy actions could conceivably affect the bond rate by altering expectations of inflation, the real rate, or both. Since policy actions may not always move the real rate and inflation components in the same direction, the near-term responses of the bond rate to such actions cannot be determined a priori. Much may depend upon the nature of the strategy being pursued by the Fed. Goodfriend (1993) discusses three different strategies that may lie at the source of interaction between the bond rate and the funds rate. Consider, first, pure cyclical strategies in which the Fed routinely raises (lowers) the funds rate in response to cyclical expansions (downturns) without attempting to affect the current trend rate of inflation expected by the public. Under that strategy, a funds rate increase will tend to raise the bond rate by raising current and expected future short real rates (i.e., by raising the rr_t^e component in [1]). This cyclical comovement is short run in nature.

The second strategy discussed by Goodfriend considers the response of the Fed to an exogenous change in the trend rate of inflation. If the trend rate of inflation increases, the bond rate automatically moves with inflation (equation

[4] The discussion in this section draws heavily from Goodfriend (1993).

[1]). The Fed may choose to keep the short real rate steady and will therefore move the funds rate in line with the rising or falling inflation rate. In this case the bond rate comoves with the funds rate because the Fed is responding to changing inflation in a neutral fashion. I refer to this source of comovement as long run in nature.

Finally, consider an aggressive strategy that could be taken either to promote real growth or to reduce the going trend rate of inflation. Under that strategy, the net impact of policy actions on the bond rate is complex because they can move the real rate (rr_t^e) and the inflation expectations ($\dot{p}_t^e$) in opposite directions. The real rate effect moves the bond rate in the same direction as the funds rate, while the inflation effect moves the bond rate in the opposite direction. Thus the net effect of policy actions on the long rate cannot be determined a priori.

To illustrate, consider an aggressive increase in the funds rate intended to reduce the trend rate of inflation. Such a tightening can shift both components of the bond rate. If the Fed's disinflation policy is credible, then short rates rise and expected inflation falls. The fall in expected inflation may thus offset somewhat the immediate response of the bond rate to the funds rate. If the decline in expected inflation persists, then the Fed may soon bring down the funds rate consistent with the lower trend rate of inflation. However, if the public does not yet have full confidence in the Fed's disinflation, then the Fed may have to persist with a sufficiently high funds rate until real growth slows and inflation actually declines. In this case, the immediate and near-term effects of the funds rate on the bond rate may be large relative to the previous case. These policy actions generate correlations between the bond rate and the funds rate which are both short and long run in nature.

Empirical Specifications of the Bond Rate Regressions: Short- and Long-Run Effects

The discussion in the previous section suggests the following observations. First, the bond rate may be correlated with current and past values of the funds rate, but the strength and duration of that correlation is a matter for empirical analysis.[5] Second, the correlation between the funds rate and the bond rate induced by pure cyclical and/or aggressive policy actions is likely to be short run, appearing over business cycle periods. In contrast, the correlation induced by the Fed's reaction to shifts in the trend rate of inflation may be long run. A rise in the trend rate of inflation that permanently raises the bond rate will

[5] This lag arises not because financial markets adjust slowly but rather because funds rate strategy puts considerable persistence in the funds rate. Such a lag can also arise if the bond rate depends upon anticipated policy moves which in turn are influenced partly by current and past values of the policy variable.

also result in a higher funds rate if the Fed is trying not to induce any cyclical or aggressive element into its policy. Third, other economic factors such as inflation, the deficit, or the state of the economy may also influence the bond rate. Such correlations are apart from the one induced by monetary policy actions.

Given the above-noted considerations, the empirical work here examines the relationship between the bond rate and the funds rate in a multivariable framework. The other economic determinants included in the analysis are the actual inflation rate and the output gap that measures the cyclical state of the economy.[6] The work identifies the short- and long-run sources of correlation between the funds rate and the bond rate using cointegration and error-correction methodology. In particular, I proceed under the assumption, whose validity I do examine, that levels of the empirical measures of the long rate, the inflation rate, the funds rate, and other economic determinants each have unit roots (stochastic trends) and that there may exist cointegrating relationships among these variables. I interpret cointegrating regressions as measuring the long-run equilibrium correlations and error-correction regressions as measuring short-run correlations among variables.

To illustrate, assume that we are examining the interaction between the bond rate and the funds rate in a system that also includes inflation. Assume further that tests for cointegration indicate the presence of the following two cointegrating regressions in the system:

$$BR_t = a_0 + a_1\dot{p}_t + U_{1t}, a_1 = 1, \text{ and} \tag{2}$$

$$NFR_t = b_0 + b_1\dot{p}_t + U_{2t}, b_1 = 1, \tag{3}$$

where BR_t is the bond rate, $\dot{p}_t$ is actual inflation, NFR is the nominal funds rate, and U_{1t} and U_{2t} are two stationary disturbances. Equation (2) indicates that the bond rate moves one-for-one with inflation and that the long real rate is mean stationary. Equation (3) indicates that the funds rate moves one-for-one with inflation and that the real funds rate is mean stationary. These two cointegrating regressions are consistent with the presence of long-run equilibrium correlations between variables. If in the cointegrating regression, say, equation (2),

[6] This framework differs somewhat from the ones used in Goodfriend (1993) and Mehra (1994). Goodfriend describes interactions between the bond rate and the funds rate, taking into account the behavior of actual inflation and real growth, whereas in Mehra (1994) the deficit also is included. That work, however, indicates that the deficit variable is not a significant determinant of the bond rate once we control for the influences of inflation and real growth. Hence the deficit variable is excluded from the work here. I use the output gap rather than real growth as a measure of the state of the economy because the bond rate appears more strongly correlated with the former than with the latter. The qualitative nature of results, however, is the same whether the output gap or real growth is used as a measure of the state of the economy. Moreover, I do examine the sensitivity of results to some changes in specification in the subsection entitled "Additional Empirical Results."

$\dot{p}_t$ is weakly exogenous, then the long-run correlation can be given a causal interpretation, implying that the bond rate is determined by the (trend) rate of inflation.[7] The hypothesis that inflation is weakly exogenous in (2) can be tested by examining whether in regressions (4) and (5)

$$\Delta BR_t = a_2 + \delta_1(BR - \dot{p})_{t-1} + \sum_{s=1}^{n1} a_{3s}\Delta BR_{t-s} + \sum_{s=1}^{n2} a_{4s}\Delta\dot{p}_{t-s} \tag{4}$$

$$\Delta\dot{p}_t = b_2 + \delta_2(BR - \dot{p})_{t-1} + \sum_{s=1}^{n1} b_{3s}\Delta BR_{t-s} + \sum_{s=1}^{n2} b_{4s}\Delta\dot{p}_{t-s}, \tag{5}$$

where $\delta_1 \neq 0$ but $\delta_2 = 0$.[8] That result indicates that it is the bond rate, not inflation, that adjusts in response to deviations in the long-run relationship.

The cointegrating regressions discussed above identify the long-run comovements among variables. In order to estimate the short-run responses of the bond rate to the funds rate, the empirical work uses the following error-correction model of the bond rate:

$$\begin{aligned}\Delta BR_t = {} & d_0 + \lambda_1 U_{1t-1} + \lambda_2 U_{2t-1} + \sum_{s=1}^{n} d_{1s}\Delta BR_{t-s} \\ & + \sum_{s=0}^{n} d_{2s}\Delta NFR_{t-s} + \sum_{s=0}^{n} d_3\Delta\dot{P}_{t-s} + \epsilon_t,\end{aligned} \tag{6}$$

where all variables are as defined before and where Δ is the first difference operator. U_{1t-1} and U_{2t-1} are one-period lagged values of the residuals from the cointegrating regressions. If we substitute for U_{1t} and U_{2t} from (2) and (3) into (6), we can rewrite (6) as in (7):

$$\begin{aligned}\Delta BR_t = {} & \tilde{d} + \lambda_1(BR_{t-1} - a_1\dot{p}_{t-1}) + \lambda_2(NFR_{t-1} - b_1\dot{p}_{t-1}) \\ & + \sum_{s=1}^{n} d_{1s}\Delta BR_{t-s} + \sum_{s=0}^{n} d_{2s}\Delta NFR_{t-s} + \sum_{s=0}^{n} d_{3s}\Delta\dot{p}_{t-s} + \epsilon_t,\end{aligned} \tag{7}$$

where $\tilde{d} = d_0 - \lambda_1 a_0 - \lambda_2 b_0$. The short-run regression (7) includes levels as well as differences of variables. The empirical effects of changes in the inflation rate and the funds rate on the bond rate may occur through two distinct channels. First, those changes may affect the bond rate directly by altering future expectations of the inflation rate and the real rate of interest. The parameters d_{is}, $i = 2,\ 3$, $s = 0,\ n$, measure near-term responses of the bond rate to changes

[7] The concept of weak exogeneity is introduced by Engle et al. (1983). The hypothesis that inflation is weakly exogenous with respect to the parameters of the cointegrating vector simply means that inferences on such parameters can be efficiently carried out without specifying the marginal distribution of inflation. More intuitively, inflation in equation (2) could be considered predetermined in analyzing the response of the bond rate to inflation.

[8] This test is proposed in Johansen (1992).

in its economic determinants. But, as noted before, signs and magnitudes of those parameters are a matter for empirical analysis because they depend upon factors such as the strategy of policy actions, the credibility of the Fed, and the nature of persistence in data. Lagged values of changes in the bond rate are included in order to capture better its own short-run dynamics.

The second focuses on disequilibrium in the long-run relations which may be caused by changes in the inflation rate and the funds rate. For example, aggressive funds rate changes taken to affect real growth or inflation may result in the level of the funds rate that is out of line with its value determined by the long-run equilibrium relation ($NFR_t \lesseqgtr b_0 + \dot{p}_t$ in [3]). Such short-run disequilibrium can also occur if the Fed adjusts the funds rate with lags in response to rising or falling inflation. Similarly, even though the bond rate moves automatically with inflation, short-run influences from other economic factors may result in the level of the bond rate that is out of line with its long-run equilibrium value ($R_t \lesseqgtr a_0 + \dot{p}_t$ in [2]). Such transitory perturbations in long-run equilibrium relations may have consequences for short-run changes in the bond rate. The parameters λ_1 and λ_2 in (7) thus measure the responses of the bond rate to such disequilibrium. The expected sign for λ_1 is negative, because the presence of the error-correction mechanism implies that the bond rate should decline (increase) if it is above (below) its long-run equilibrium value. In contrast, the sign of λ_2 is expected to be positive. But note all these disequilibrium effects are short-run (cyclical) in nature because in the long run (defined here in the equilibrium sense) they disappear and the bond rate is at its long-run equilibrium value determined by (2), i.e., $a_0 + \dot{p}_t$.

Data and Estimation Procedures

The empirical work in this article focuses on the behavior of the long rate during the sample period from 1957Q1 to 1995Q2. The long rate is measured by the nominal yield on ten-year U.S. Treasury bonds (*BR*). In most previous studies a distributed lag on the actual inflation rate is used as proxy for the long-run anticipated inflation, and actual inflation is generally measured by the behavior of the consumer price index. I also use actual inflation as proxy for anticipated inflation. I, however, measure inflation as the average of change in the consumer price index, excluding food and energy, over the past three years ($\dot{p}$).[9] The output gap (gaph) is the natural log of real GDP minus the

[9] I get similar results if instead the consumer price index or the GDP deflator is used to measure actual inflation (see the subsection entitled "Additional Empirical Results").

In a couple of recent studies (Hoelscher 1986; Mehra 1994) the Livingston survey data on one-year-ahead inflationary expectations are used to measure long-run anticipated inflation. The results in Mehra (1994), however, indicate that the near-term impact of the funds rate on the bond rate remains significant if one-year-ahead expected inflation (Livingston) data are substituted for actual inflation in the empirical work (see Mehra 1994, Table 4). That result continues to hold in this article also (see the subsection entitled "Additional Empirical Results").

log of potential GDP, which is generated using the Hodrick-Prescott filter. The interest rates are monthly averages for the last month of the quarter.

The stationarity properties of the data are examined using tests for unit root and mean stationarity. The unit root test used is the augmented Dickey-Fuller test and the test for mean stationarity is the one in Kwiatkowski et al. (1992). The test used for cointegration is the one proposed in Johansen and Juselius (1990).[10]

2. EMPIRICAL FINDINGS

In this section I describe cointegration test results for a system that includes the bond rate (*BR*), the inflation rate ($\dot{p}$), and the nominal funds rate (*NFR*). I also discuss short-run results from error-correction regressions for the full sample period as well as for several subperiods. The section concludes with an explanation of different pre- and post-1979 sample results.

Cointegration Test Results

Test results for unit roots and mean stationarity are summarized in the appendix. They indicate that the bond rate (*BR*), the inflation rate ($\dot{p}$), and the nominal funds rate (*NFR*) each have a unit root and thus contain stochastic trends. The output gap variable by construction is stationary.

Test results for cointegration are also summarized in the appendix. I first focus on the bivariable systems ($BR, \dot{p}$), ($NFR, \dot{p}$), and (BR, NFR). Test results are consistent with the presence of cointegration between variables in each system, indicating that the bond rate is cointegrated with the inflation rate and the funds rate. The funds rate is also cointegrated with the inflation rate. Thus the bond rate comoves with each of these nonstationary variables, including the funds rate.

The presence of cointegration between two variables simply means that there exists a long-run stochastic correlation between them. In order to help determine whether such correlation can be given a causal interpretation, Table 1 presents test results for weak exogeneity of the long-run parameters. In the system ($BR, \dot{p}$) inflation is weakly exogenous but the bond rate is not, indicating that it is the bond rate that adjusts in response to deviations in the long-run relationship. Thus the long-run equilibrium relationship between the bond rate and the inflation rate can be interpreted as a Fisher relation in which the bond rate is determined by the (trend) rate of inflation. In the system ($NFR, \dot{p}$) inflation is again weakly exogenous but the funds rate is not. Here again the long-run relation can be interpreted as one in which the inflation rate drives the interest rate: in this case the short-term rate. Hence, I interpret the long-run

[10] These tests are described in Mehra (1994).

Table 1 Cointegrating Regressions and Test Results for Weak Exogeneity

Equation Number	Panel A Cointegrating Regressions	Panel B Error-Correction Coefficients (t-value) in Equations for		
		ΔBR	$\Delta \dot{p}$	ΔNFR
1	$BR_t = 3.3 + 0.93\ \dot{p}_t + U_{1t}$ (10.0)	−0.18 (3.4)	−0.01 (0.6)	
2	$BR_t = 1.2 + 0.91\ NFR_t + U_{2t}$ (50.7)	−0.17 (2.5)		0.21 (1.5)
3	$NFR_t = 1.8 + 1.1\ \dot{p}_t + U_{3t}$ (6.2)		0.00 (0.4)	−0.27 (3.2)
4	$(BR - \dot{p})_t = 2.2 + 0.09\ NFR_t$ (1.3)			
	$\chi_1^2 = 1.7$			

Notes: Cointegrating regressions given in panel A above are estimated by the dynamic OLS procedure given in Stock and Watson (1993), using leads and lags of first differences of the relevant right-hand side explanatory variables. Eight leads and lags are included. Parentheses that appear below coefficients in cointegrating regressions contain t-values corrected for the presence of moving average serial correlation. The order of serial correlation was determined by examining the autocorrelation function of the residuals. χ_1^2 is the Chi-square statistic that tests the hypothesis that the coefficient that appears on *NFR* in equation 4 is zero.

Panel B above contains error-correction coefficients from regressions of the form

$$\Delta X_{1t} = \delta_1\ U_{t-1} + \sum_{s=1}^{4} a_s \Delta X_{1t-s} + \sum_{s=1}^{4} b_s\ \Delta X_{2t-s}$$

$$\Delta X_{2t} = \delta_2\ U_{t-1} + \sum_{s=1}^{4} c_s \Delta X_{1t-s} + \sum_{s=1}^{4} d_s\ \Delta X_{2t-s},$$

where U_{t-1} is the lagged value of the residual from the cointegrating regression that is of the form

$$X_{1t} = d_0 + d_1 X_{2t} + U_t,$$

and where X_{1t} and X_{2t} are the pertinent nonstationary variables. The relevant cointegrating regressions are given in panel A above. Parentheses that appear below error-correction coefficients contain t-values.

equilibrium relationship between the funds rate and the inflation rate as a kind of reaction function.

The test results for weak exogeneity discussed above for systems $(BR, \dot{p})$ and $(NFR, \dot{p})$ also imply that inflation causes the comovement between the bond rate and funds rate. The bond rate comoves with the funds rate because the bond rate moves automatically with inflation and the Fed keeps the funds rate in line with the trend rate of inflation in the long run.

The analysis above, based on bivariable systems, thus suggests that in the Fisher relation the funds rate should not be correlated with the bond rate once we control for the correlation that is due to inflation. I test this implication by examining whether the ex post real rate ($BR - \dot{p}$) is correlated with the funds rate. I do so by expanding the Fisher relation to include the funds rate while maintaining the Fisher restriction that the bond rate adjusts one-for-one with inflation. In that regression the funds rate is not significant (see Table 1, equation 4). The Chi-square statistic for the null hypothesis that the ex post real rate is not correlated with the funds rate is small, consistent with no correlation.

The result that the bond rate is cointegrated with the actual inflation rate and thus the ex post real rate ($R_t - \dot{p}_t$) is stationary also implies that the expected long real rate is stationary. This can be seen if we express the Fisher relation (1) as

$$BR_t - \dot{p}_t = rr_t^e + (\dot{p}_t^e - \dot{p}_t) = rr_t^e + U_t,$$

where all variables are defined as before and where U_t is the disturbance term. This disturbance arises because the long-term expected inflation rate may differ from the three-year inflation rate. As is clear, the stationarity of ($BR_t - \dot{p}_t$) implies the stationarity of the expected long real rate.

Error-Correction Regressions

Since the ex ante long real rate is mean stationary, not constant, cyclical and aggressive funds rate changes discussed before may still affect the bond rate by altering expectations of its real rate and inflation components. I now explore those short-run effects by estimating the error-correction equation.

The empirical results discussed in the previous section are consistent with the following two cointegrating relationships:

$$BR_t = a_0 + a_1\dot{p}_t + U_{1t} \quad \text{and} \tag{8}$$

$$NFR_t = b_0 + b_1\dot{p}_t + U_{2t}. \tag{9}$$

Equation (8) is the Fisher relation, and equation (9) the Fed reaction function. The latter captures that component of the funds rate that comoves with trend inflation. The residual U_{2t} is then the component that captures the stance of cyclical and aggressive funds rate policy actions. Consider then the following error-correction equation:

$$\Delta BR_t = d_0 + \lambda_1 U_{1t-1} + \lambda_2 U_{2t-1} + \sum_{s=1}^{n} d_{1s}\Delta BR_{t-s} + \sum_{s=1}^{n} d_{2s}\Delta\dot{p}_{t-s} + \sum_{s=1}^{n} d_{3s}NFR_{t-s} + \sum_{s=1}^{n} d_{4s}gaph_{t-s} + \epsilon_t. \tag{10}$$

The parameter λ_2 in (10) measures the response of the bond rate to the lagged value of the funds rate spread ($NFR - b_0 - b_1\dot{p}$). Since the short-run equation is in first differences, the near-term response of the bond rate to the funds rate spread can be calculated as $-\lambda_2/\lambda_1$.[11] Also, equation (10) includes only lagged values of economic determinants and hence can be estimated by ordinary least squares. In order to examine subsample variability, I estimate equations (8) through (10) over several sample periods, all of which begin in 1961Q2 but end in different years from 1972 through 1995.

Table 2 presents some key coefficients ($\lambda_1, \lambda_2, -\lambda_2/\lambda_1, a_1, b_1$) from these regressions. If we focus on full sample results, then it can be seen that all these key coefficients appear with expected signs and are statistically significant. In cointegrating regressions the bond rate adjusts one-for-one with inflation and so does the funds rate. The hypotheses that $a_1 = 1$ and $b_1 = 1$ cannot be rejected and thus are consistent with our priors about the interpretation of (8) as the Fisher relation and of (9) as the Fed reaction function. In the error-correction regression λ_2 is positive and its estimated value indicates a one percentage point rise in the funds rate spread raises the bond rate by 13 basis points in the following period. The net increase totals 42 basis points. The mean lag ($-1/\lambda_1$) in the short-run effect of the funds rate spread on the bond rate is approximately 3.2 quarters, indicating that these near-term responses dissipate quite rapidly.[12]

---

[11] This can be shown as follows. Assume, for example, the level of the bond rate is related to inflation and the funds rate spread as in

$$BR_t = a_0 + a_1\dot{p}_t + a_2(NFR - b_1\dot{p})_t + V_t, \tag{a}$$

where the stationary component is $a_0 + a_2(NFR - b_1\dot{p})_t$ and the nonstationary component is $a_1\dot{p}_t$. The error-correction regression is

$$\Delta BR_t = \lambda_1 V_{t-1} + \sum_{s=1}^{n} \text{(other lagged differences of variables)} + \epsilon_t, \tag{b}$$

where λ_1 is negative. Substituting for V_{t-1} from (a) into (b) yields (c):

$$\Delta BR_t = \lambda_1 BR_{t-1} - \lambda_1 a_1 \dot{p}_{t-1} - \lambda_1 a_2 (NFR - b_1\dot{p})_{t-1} + \textit{ other terms}. \tag{c}$$

Equation (c) can be estimated and a_2 can be recovered as $\lambda_1 a_2/\lambda_1$, which is the minus of the coefficient on $(NFR - b_1\dot{p})_{t-1}$ divided by the coefficient on BR_{t-1}. The coefficient a_2 then measures the near-term response of the bond rate to the funds rate spread. I do not label a_2 as measuring the long-run effect because the spread is stationary. The long run is defined as the period over which trend relationships emerge. In the long run the funds rate spread ($NFR - \dot{p}$) is constant.

[12] In estimated short-run regressions the coefficients that appear on lagged differences of the bond rate are very small. If we ignore those coefficients, then the short-run equation (c) given in footnote 11 can be expressed as

$$BR_t = \frac{-\lambda_1 a_1}{1 - (1 + \lambda_1)L}\,\dot{p}_{t-1} + \frac{\lambda_2}{1 - (1 + \lambda_1)L}\,(NFR - \dot{p})_{t-1} \; + \; \textit{other terms},$$

where L is the lag operator and where λ_2 is $-\lambda_1 a_2$. The coefficients (w_i) that appear on lagged levels of $NFR - \dot{p}$ are then of the form $\lambda_2, \lambda_2(1 + \lambda_1), \lambda_2(1 + \lambda_1)^2$, etc. The mean lag then can be calculated as follows:

If we focus on subsample results, it can be seen that all key coefficients still appear with expected signs and are statistically different from zero. However, there are some major differences between pre- and post-1979 regression estimates. In pre-1979 cointegrating regressions the hypotheses that $a_1 = 1$ and $b_1 = 1$ are generally rejected. In contrast, that is not the case in most post-1979 regressions. The rejection, however, is more common in the Fisher relation than it is in the Fed reaction function.

In pre-1979 error-correction regressions the bond rate does respond to the funds rate spread—the one-period response (λ_2) ranges from 11 to 23 basis points. But the one-period response is generally quite close to the near-term net response ($-\lambda_2/\lambda_1$), indicating that the effect of policy actions on the bond rate did not persist too long. In post-1979 error-correction regressions, however, the near-term response of the bond rate to the funds rate spread is larger than the one-period lagged response. In particular, the immediate response of the bond rate to the funds rate ranges from 13 to 16 basis points and the near-term response from 36 to 48 basis points. Together these estimates imply that the near-term response of the bond rate to the funds rate spread has increased since 1979.[13] I argue below that these different results may be due in part to the way the Fed has conducted its monetary policy since 1979. In particular, I focus on the Fed's disinflation policy.

Before 1979 the Fed did not aggressively attempt to bring down the trend rate of inflation. In the long-run Fed reaction function estimated over 1961Q2 to 1979Q3, the parameter b_1 is less than unity, indicating that the Fed did not adjust the funds rate one-for-one with inflation (see Table 2). Moreover, the short-run reaction functions estimated in Mehra (1996) also indicate that in the pre-1979 period the Fed did not respond to accelerations in actual inflation. Hence, during this early period a monetary policy tightening measured by a widening in the funds rate spread may have affected the bond rate primarily by altering its expected real rate component. Because the funds rate increase alters only near-term expectations of future short real rates, its impact on the bond rate is likely to be modest, as confirmed by low estimates of λ_2 in Table 2. When the bond rate rises above the current inflation rate, both the bond rate and actual inflation rises, speeding up the adjustment, as confirmed by high estimates of λ_1 in Table 2. As a result, the immediate effect of policy on

$$\textit{Mean Lag} = \sum_{i=1}^{\infty} w_i\, i \,/ \sum_{i=1}^{\infty} w_i = \frac{\lambda_2[1 + 2(1+\lambda_1) + 3(1+\lambda_1)^2 + 4(1+\lambda_1)^3 + \ldots \infty]}{\lambda_2[1 + (1+\lambda_1)^2 + (1+\lambda_1)^3 + \ldots \infty]}$$
$$= \lambda_2 \frac{1}{(1-1-\lambda_1)^2} \times \frac{1}{\lambda_2 \frac{1}{1-1-\lambda_1}} = -\frac{\lambda_1}{\lambda_1^2} = -\frac{1}{\lambda_1}.$$

[13] This is confirmed by results of the formal Chow test that is discussed in the next subsection.

Table 2 Short-Run Error-Correction Regressions Using Residuals

Estimation Ends in Year	Panel A: Error-Correction Coefficients				Panel B: Coefficients from Cointegrating Regressions	
	$\lambda_2(\tilde{t})$	$\lambda_1(\tilde{t})$	$-\lambda_2/\lambda_1$	Mean Lag in Quarters	$a_1(\tilde{t})$	$b_1(\tilde{t})$
1995Q2	0.13(2.7)	−0.31(4.8)	0.42	3.2	0.93(10.0)	1.10 (6.0)
1994	0.12(2.4)	−0.29(4.5)	0.41	3.4	0.93(10.1)	1.10 (6.2)
1992	0.14(2.6)	−0.30(4.5)	0.46	3.3	0.92 (9.7)	1.10 (6.5)
1990	0.15(2.6)	−0.31(4.3)	0.47	3.3	0.92 (9.8)	1.10 (6.7)
1988	0.14(2.4)	−0.30(4.1)	0.48	3.3	0.94(10.9)	1.00 (6.6)
1986	0.16(2.7)	−0.34(2.7)	0.48	2.9	0.94(10.8)	1.10 (6.9)
1984	0.13(1.9)	−0.30(2.6)	0.46	3.3	0.93(11.1)	1.10 (6.7)
1982	0.19(2.5)	−0.54(3.6)	0.36	1.8	0.76(17.1)[c]	0.84 (7.6)
1980	0.18(2.1)	−0.71(2.2)	0.26	1.4	0.68(30.8)[c]	0.60 (5.6)[c]
1979Q3	0.17(3.1)	−0.67(3.3)	0.25	1.5	0.69(31.1)[c]	0.77 (8.9)[c]
1978	0.16(2.8)	−0.64(3.1)	0.26	1.5	0.71(28.4)[c]	0.78 (9.9)[c]
1976	0.17(3.2)	−0.65(3.0)	0.26	1.5	0.71(21.4)[c]	1.00(15.9)
1974	0.23(2.9)	−0.98(3.6)	0.23	1.0	0.78(17.9)[c]	0.97(14.6)
1972	0.11(1.2)	−0.75(2.6)	0.14	1.3	0.73 (6.9)[c]	0.68 (2.5)

[c] indicates the relevant coefficient (a_1 or b_1) is significantly different from unity.

Notes: The coefficients reported above are from the following regressions:

$$BR_t = a_0 + a_1\dot{p}_t + U_{1t}, \tag{a}$$

$$NFR_t = b_0 + b_1\dot{p}_t + U_{2t}, \text{ and} \tag{b}$$

$$\Delta BR_t = d_0 + \lambda_1 U_{1t-1} + \lambda_2 U_{2t-1} + \sum_{s=1}^{n} d_{1s}\Delta BR_{t-s} + \sum_{s=1}^{2} d_{2s}\Delta\dot{p}_{t-s} + \sum_{s=1}^{2} d_{3s}\Delta NFR_{t-s} + \sum_{s=1}^{2} d_{4s}\, gaph_{t-s} + \epsilon_t, \tag{c}$$

where gaph is the output gap and other variables are defined as in Table A1. Equations (a) and (b) are estimated by the dynamic OLS procedure given in Stock and Watson (1993) and equation (c) by ordinary least squares. The estimation period begins in 1961Q2 and ends as shown in the first column. $\tilde{t}$ is the t-statistic. The mean lag is calculated as $-1/\lambda_1$.

the bond rate dissipates quickly and hence the near-term effect is close to the immediate impact.

In the post-1979 period, however, the Fed made a serious attempt to bring down the trend rate of inflation and to contain inflationary expectations. The descriptive analysis of monetary policy in Goodfriend (1993) and the short-run reaction function estimated in Mehra (1996) are consistent with this observation. Hence short-run increases in the funds rate spread may also have

affected the bond rate by altering the long-term expected inflation. If the Fed's disinflation policy had been credible, then increases in the funds rate spread that raise the bond rate's real component may also lower expectations of the long-term expected inflation rate, thereby offsetting somewhat the immediate or the very-near-term response of the bond rate to the funds rate spread. The evidence reported in Table 2, however, indicates that the estimated coefficient $(-\lambda_2/\lambda_1)$ that measures the near-term response shows no tendency to fall in the post-1979 period. On the other hand, if the public does not have full confidence in the Fed's disinflation policy and if the Fed has to persist with sufficiently high short real rates to reduce the trend rate of inflation or contain inflationary expectations, then the estimated effect of a policy action on the bond rate would last longer. In this case, the near-term effect of the funds rate spread on the bond rate would be larger and the mean lag in the effect of such policy on the bond rate would also be higher. Both these implications are consistent with results in Table 2, where both the estimated short-run effect $(-\lambda_2/\lambda_1)$ and the mean lag $(-1/\lambda_1)$ are higher in the post-1979 period than they were in the period before.

Additional Empirical Results

In this section I discuss and report some additional test results which confirm the robustness of conclusions reached in the previous section. I consider several changes in the specification of the short-run equation (10). First, I reproduce the empirical work in Table 2 under the alternative specification that the short-run stance of monetary policy is measured by the real funds rate. I then consider additional changes in specification to address concerns raised by the potential endogeneity of monetary policy actions, alternative measures of inflation, and the potential stationarity of data. For these latter experiments I focus on results that pertain to the short-run effect of the funds rate on the bond rate over two sample periods only, 1961Q1 to 1995Q2 and 1961Q2 to 1979Q3. Hence I report two key coefficients, λ_2 and $(-\lambda_2/\lambda_1)$.

The results in Table 2 discussed in the previous section use the residual from the long-run Fed reaction function as a measure of the short-run stance of monetary policy. I now examine results if the short-run stance of policy is measured instead by the real funds rate $(NFR-\dot{p})$. Furthermore, I now estimate the Fisher relation (8) jointly with the short-run error-correction equation. This procedure allows for richer short-run dynamics in estimating the long-run effect of inflation on the bond rate. The short-run equation that incorporates these two new changes can be derived by replacing the residuals U_{1t-1} and U_{2t-1} in (10) by lagged levels of the variables and then by setting $b_1 = 1$. The resulting short-run equation is

$$\Delta BR_t = \tilde{d} + \lambda_1 BR_{t-1} + \lambda_3 \dot{P}_{t-1} + \lambda_2 (NFR - \dot{p})_{t-1} + \sum_{s=1}^{n} d_{1s} \Delta BR_{t-s} + \sum_{s=1}^{n} d_{2s} \Delta \dot{p}_{t-s} + \sum_{s=1}^{n} d_{3s} \Delta (NFR - \dot{p})_{t-s} + \sum_{s=1}^{n} d_{4s}\ gaph_{t-s} + \epsilon_t, \quad (11)$$

where $\lambda_3 = -\lambda_1 a_1$. In (11), if $a_1 = 1$, then the coefficients on BR_{t-1} and $\dot{p}_{t-1}$ sum to zero ($\lambda_1 + \lambda_3 = 0$ in (11)). I impose this restriction only if it is not rejected. Table 3 reports some key coefficients ($\lambda_1, \lambda_2, -\lambda_2/\lambda_1, \lambda_3$). These estimated coefficients confirm the qualitative nature of results in Table 2. First, the real funds rate is a significant predictor of the bond rate and this result

Table 3 Short-Run Error-Correction Regressions Using the Level of the Funds Rate Spread

Estimation Ends in Year	$\lambda_2(\tilde{t})$	$\lambda_1(\tilde{t})$	$-\lambda_2/\lambda_1$	Mean Lag in Quarters	$\lambda_3(\tilde{t})$	$-\lambda_3/\lambda_1$	F1
1995Q2	0.12(2.6)	−0.29(4.7)	0.40	3.5	0.29(4.7)	1.00	0.89
1994	0.10(2.2)	−0.26(4.3)	0.35	3.8	0.26(4.3)	1.00	0.89
1992	0.11(2.4)	−0.28(4.4)	0.39	3.6	0.28(4.4)	1.00	0.64
1990	0.11(2.2)	−0.28(4.1)	0.39	3.6	0.28(4.1)	1.00	0.68
1988	0.11(2.2)	−0.27(3.9)	0.41	3.7	0.27(3.9)	1.00	0.63
1986	0.13(2.5)	−0.30(4.2)	0.41	3.2	0.30(4.2)	1.00	0.55
1984	0.10(1.7)	−0.22(2.3)	0.41	4.5	0.22(2.3)	1.00	0.11
1982	0.21(3.4)	−0.53(3.5)	0.40	1.9	0.48(3.5)	0.90	3.70*
1980	0.12(1.9)	−0.24(2.0)	0.50	4.2	0.24(2.0)	1.00	1.00
1979Q3	0.13(2.5)	−0.41(2.4)	0.31	2.4	0.32(2.5)	0.80	3.60*
1978	0.12(2.4)	−0.41(2.3)	0.29	2.4	0.32(2.3)	0.78	3.40*
1976	0.14(2.6)	−0.53(2.6)	0.26	1.9	0.39(2.5)	0.74	6.60**
1974	0.22(3.1)	−0.92(3.8)	0.23	1.1	0.67(3.7)	0.73	14.20**
1972	0.21(2.3)	−0.97(3.4)	0.22	1.0	0.69(3.3)	0.71	10.10**

*Significant at the 10 percent level.

**Significant at the 5 percent level.

Notes: The coefficients reported are from regressions of the form

$$\Delta BR_t = d_0 + \lambda_1\ BR_{t-1} + \lambda_3\ \dot{P}_{t-1} + \lambda_2\ (NFR - \dot{p})_{t-1} + \sum_{s=1}^{2} d_{1s}\ \Delta BR_{t-s} + \sum_{s=1}^{2} d_{2s}\ \Delta \dot{p}_{t-s} + \sum_{s=1}^{2} d_{3s}\ \Delta (NFR - \dot{p})_{t-s} + \sum_{s=1}^{2} d_{4s}\ gaph_{t-s},$$

where gaph is the output gap and other variables are as defined in Table A1. All regressions are estimated by ordinary least squares. The estimation period begins in 1961Q2 and ends in the year shown. The mean lag is calculated as $-1/\lambda_1$. $\tilde{t}$ is the t-statistic. F1 tests the null hypotheses that λ_1 and λ_3 sum to zero, indicating that the Fisher restriction is consistent with data.

is fairly robust over several subsamples. Second, the Chow test indicates that two key parameters, λ_1 and λ_2, are unstable only between pre- and post-1979 periods (the date of the break is 1980Q2). This result is consistent with an increase in the near-term effect of policy on the bond rate in the post-1979 period.

The short-run error-correction equation (11) was alternatively re-estimated using the consumer price index and the GDP deflator to measure inflation, that is, the average inflation rate over the past three years. As in a few previous studies, I also used the Livingston survey data on one-year-ahead inflationary expectations. The results continue to indicate that the funds rate spread generally does help predict the bond rate and that the near-term response of the bond rate to the funds rate spread has increased since 1979 (see rows 1 through 3 in Table 4).[14]

The funds rate spread here measures the short-run stance of monetary policy because it is that component of the funds rate that does not comove with inflation. This spread, however, is still endogenous because, as noted before, the Fed routinely raises the funds rate during cyclical expansions and lowers it during cyclical downturns. The potential problem created by such endogeneity is that if the bond rate is directly influenced by variables that reflect the cyclical state of the economy and if those variables are omitted from short-run regressions, then the funds rate spread may be picking up the influence of those variables on the bond rate rather than the influence of monetary policy on the real component of the bond rate.

The short-run regressions reported in Tables 2 and 3 already include many of those variables such as the output gap that measures the cyclical state of the economy and changes in inflation, the bond rate, and the funds rate spread itself. While it is difficult to know all the information that the Fed may be using in setting its short-run funds rate policy, I re-estimated (11) alternatively including additional variables such as nonfarm payroll employment, sensitive materials prices, the deficit, and real growth. Those additional variables, when included in (11), are generally not significant and therefore do not change the qualitative nature of results in Table 3 (see rows 4a through 4d in Table 4).

It is sometimes argued that unit root tests used here have low power in distinguishing whether the variables are stationary or integrated. Hence the cointegration and error-correction methodology used here to distinguish between long- and short-run sources of comovement between the bond rate and the funds rate is suspect. However, the evidence presented above that the bond rate and the funds rate adjust one-for-one with inflation in the long run is confirmed even if I treat the bond rate, the funds rate, and inflation as stationary

[14] When inflation is measured by the behavior of the consumer price index or the Livingston survey, I get some mixed results. The statistical significance of the coefficient that appears on the funds rate spread is not robust over different sample periods.

Table 4 Sensitivity to Changes in Specification

Changes in Specification	Panel A: 1961Q2–1995Q2				Panel B: 1961Q2–1979Q3			
	$\lambda_2(\tilde{t})$	$(-\lambda_2/\lambda_1)$	f_2	$\tilde{f}_2$	$\lambda_2(\tilde{t})$	$(-\lambda_2/\lambda_1)$	f_2	$\tilde{f}_2$
1. CPI	0.08(1.7)	0.42			0.09(1.6)	0.16		
2. GDP Deflator	0.11(2.2)	0.49			0.10(2.0)	0.19		
3. Livingston Survey	0.08(1.7)	0.35			0.07(1.6)	0.23		
4. CPIEXFE Additional Variables								
a. $\Delta \ln PEM$	0.14(2.9)	0.48			0.15(2.8)	0.35		
b. d_t	0.10(2.3)	0.37			0.13(2.0)	0.28		
c. $\Delta \ln SMP$	0.10(2.4)	0.38			0.13(2.5)	0.29		
d. $\Delta \ln ry_t$	0.12(2.5)	0.43			0.10(1.5)	0.25		
5. Stationary: Level Regressions			0.07(1.9)	0.42			0.09(2.1)	0.29

Notes: The coefficients reported are from regressions of the form given in Table 3. The Fisher restriction is imposed in regressions estimated over 1961Q2–1995Q2 but not in those estimated over 1961Q2 to 1979Q3. CPI is the consumer price index; CPIEXFE is the consumer price index excluding food and energy; PEM is the nonfarm payroll employment; $\Delta \ln ry$ is real GDP growth; d is federal government deficits scaled by nominal GDP, and SMP is the sensitive materials prices. The coefficients reported in the row labeled 5 are from the regression (14) of the text. f_2 is the coefficient that measures the contemporary response of the bond rate to the funds rate spread and $\tilde{f}_2$ the near-term.

variables. The stationary versions of the long-run regressions (8) and (9) can be expressed as in (12) and (13):

$$BR_t = a_0 + \sum_{s=0}^{n} a_{1s}\ \dot{p}_{t-s} + \sum_{s=1}^{n} a_{2s}\ BR_{t-s} + U_{1t} \tag{12}$$

$$NFR_t = b_0 + \sum_{s=0}^{n} b_{1s}\ \dot{p}_{t-s} + \sum_{s=1}^{n} b_{2s}\ NFR_{t-s} + U_{2t}. \tag{13}$$

The net response of the bond rate to inflation is $\left(\sum_{s=0}^{n} a_{1s} \Big/ 1 - \sum_{s=1}^{n} a_{2s}\right)$ and to the funds rate is $\left(\sum_{s=0}^{n} b_{1s} \Big/ 1 - \sum_{s=1}^{n} b_{2s}\right)$. One cannot reject the hypotheses that these net responses each are unity. As for short-run correlations, consider the following stationary version of the short-run equation:

$$BR_t = f_0 + f_1 \dot{p}_t + f_2\ (NFR - \dot{p})_t + f_3\ gaph_t + \sum_{s=1}^{n}\ f_{4s}\ BR_{t-s} + \epsilon_t, \tag{14}$$

where all variables are as defined before. Equation (14) already incorporates the long-run restriction that short-run funds rate policy actions affect the bond rate by altering the spread between the funds rate and the inflation rate. The other restriction can be imposed by the requirement that coefficients f_1 and $\sum_{s=1}^{n} f_{4s}$ in (14) sum to unity. The parameter f_2 in (14) measures the contemporaneous response of the bond rate to the funds rate spread and its net response can be calculated as $\left(f_2 / 1 - \sum_{s=1}^{n} f_{4s}\right)$.

I estimate equation (14) by instrumental variables that are just lagged values of the right-hand side explanatory variables. In such regressions the funds rate spread still helps predict the bond rate (see row 5 in Table 4).

A Comparison with Some Previous Studies

In this section I discuss some previous studies that use entirely different methodologies but reach conclusions regarding the short-run impact of policy on long-term rates which are qualitatively similar to those reported here.

The first set consists of studies by Cook and Hahn (1989), Radecki and Reinhart (1994), and Roley and Sellon (1995). All three of these studies examine the response of long-term interest rates to changes in a measure of the federal funds rate target. They differ, however, with respect to the sample period studied and the length of the interval over which the interest rate response is measured. In Cook and Hahn the sample period studied is September 1974 to September 1979 and the interest rate response is measured on the day of the target change. In the other two studies the sample periods examined fall

within the post-1979 period: 1989 to 1993 in Radecki and Reinhart (1994) and October 1987 to July 1995 in Roley and Sellon (1995). The measurement interval in Radecki and Reinhart spans the first ten days following the policy change, whereas in Roley and Sellon the time interval spans the period from the day after the previous policy action to the day after the current policy action. The economic rationale for the use of a wider measurement interval as in Roley and Sellon is that many times monetary policy actions are already anticipated by the markets, so that long-term interest rates move ahead of the announced change in the funds rate target. The relative magnitudes of interest rate responses before, during, and after the policy action, however, depend upon whether policy actions are anticipated or unanticipated and upon the degree of persistence in anticipated policy actions.

The measurement interval is the narrowest in Cook and Hahn and Radecki and Reinhart; the results there indicate that a one percentage point increase in the funds rate target induces 12 to 13 basis points movement in the ten-year bond rate (increases in longer maturity bond rates are somewhat smaller). The size of the interest rate response during and after the change in policy action as measured by Roley and Sellon is also modest; the 30-year Treasury bond yield rises by 10 basis points following one percentage point increase in the effective federal funds rate target. However, when the measurement interval includes days before the change in policy action, the measured interest rate response rises to 38 basis points. Thus a significant part of the response occurs before policy action is announced, indicating the presence of anticipated effect. What needs to be noted is that the magnitude of the total interest rate response measured by Roley and Sellon is quite close to the near-term response that I have estimated using an entirely different estimation methodology. Recall that for the complete sample period 1961Q2 to 1995Q2 the estimated near-term response of the ten-year bond rate to the funds rate spread is 42 basis points (see Table 2). The estimated near-term response is 36 basis points if instead I use the 30-year bond yield in my empirical work.

The other recent study showing that in the short run the long real rate comoves with the short nominal rate is by Fuhrer and Moore (1995). According to the expectational theory of the term structure of interest rates, the ex ante long-term real rate can be viewed as a weighted moving average of future short real rates. Fuhrer and Moore define the short real rate as the nominal yield on three-month Treasury bills minus the actual quarterly inflation rate. They then use a vector autoregression to construct long-horizon forecasts of the time paths of the three-month Treasury bill rate and the inflation rate. Given those forecasts they compute the 40-quarter-duration long-term real rate, since the average duration (maturity) of Moody's BAA corporate bond rate is 40 quarters. What they find is that over the period 1965 to 1992 the sample path of the ex ante long real rate tracks closely that of the short-term nominal rate (Fuhrer and Moore 1995, Figure 1, p. 224). The ex ante long real rate is still

relatively stable, however, and only about one-fourth of the increase in the short nominal rate is reflected in the long real rate.

3. CONCLUDING OBSERVATIONS

This article has investigated empirically the immediate, near-term, and long-run effects of monetary policy on the bond rate. The federal funds rate is used as a measure of monetary policy, and the long run is viewed as the period during which trend relationships emerge. The results indicate that the long-run effect of monetary policy on the bond rate occurs primarily through the inflation channel.

In the short run, however, monetary policy also affects the bond rate by altering its expected real rate component. The short-run stance of monetary policy is measured by the spread between the funds rate and the ongoing trend rate of inflation. The results indicate that the near-term effect of the funds rate spread on the bond rate has increased considerably since 1979. In the pre-1979 period, the bond rate rose anywhere from 14 to 29 basis points whenever the funds rate spread widened by one percentage point. In the post-1979 period, however, the estimate of its near-term response ranges from 26 to 50 basis points.

This increase in the short-run sensitivity of the bond rate to monetary policy actions is consistent with the way the Fed has conducted its monetary policy since 1979. Since then the Fed has made a serious attempt to bring down the trend rate of inflation and contain inflationary expectations. If the public does not have full confidence in the Fed's disinflation policy, and if the Fed has to persist with sufficiently high short real rates to reduce the trend rate of inflation or contain inflationary expectations, then the estimated effect of a policy action on the bond rate would last longer. As a result, the near-term effect of policy on the bond rate would be stronger than would be the case if the disinflation policy were fully credible.

APPENDIX A

The stationarity properties of data are investigated using both unit roots and mean stationarity tests. The test for unit roots used is the augmented Dickey-Fuller test and the one for mean stationarity is the one in Kwiatkowski et al. (1992). Both these tests are described in Mehra (1994).

Table A1 presents test results for determining whether the variables BR, $\dot{p}$, NFR, $BR - \dot{p}$ and $NFR - \dot{p}$ have a unit root or are mean stationary. As can be

Table A1 Tests for Unit Roots and Mean Stationarity

Series	Panel A Tests for Unit Roots			Panel B Tests for Mean Stationarity
X	$\hat{\rho}$	$t_{\hat{\rho}}$	k	$\hat{n}_u$
BR	0.96	−1.7	5	0.83*
$\dot{p}$	0.99	−2.0	1	0.56*
NFR	0.89	−2.9[a]	5	0.46[a]
$BR - \dot{p}$	0.87	−3.2*	3	0.04
$NFR - \dot{p}$	0.70	−5.0*	5	0.03
ΔBR	−0.10	−5.6*	4	0.19
$\Delta\dot{p}$	0.60	−4.3*	7	0.18
ΔNFR	−0.30	−4.9*	7	0.07

[a]The test statistic is close to the relevant 5 percent critical value.

*Significant at the 5 percent level.

Notes: BR is the ten-year Treasury bond rate; $\dot{p}$ is the average inflation rate over the past three years; and NFR is the nominal funds rate. Δ is the first difference operator. Inflation is measured by the behavior of the consumer price index excluding food and energy. This price series begins in 1957; therefore the effective sample period studied is 1960Q1 to 1995Q2. The values for ρ and t-statistics ($t_{\hat{\rho}}$) for $\rho = 1$ in panel A above are from the augmented Dickey-Fuller regressions of the form

$$X_t = a_0 + \rho X_{t-1} + \sum_{s=1}^{k} a_s \, \Delta X_{t-s}, \tag{a}$$

where X is the pertinent series. The number of lagged first differences (k) included in these regressions are chosen using the procedure given in Hall (1990). The procedure starts with some upper bound on k, say k max chosen a priori (eight quarters here). Estimate (a) above with k set at k max. If the last included lag is significant, select $k = k$ max. If not, reduce the order of the autoregression by one until the coefficient on the last included lag is significant. The test statistic $\hat{n}_u$ in panel B above is the statistic that tests the null hypothesis that the pertinent series is mean stationary. The 5 percent critical value for $\hat{n}_u$ given in Kwiatkowski et al. (1992) is 0.463 and for $t_{\hat{\rho}}$ given in Fuller (1976) is −2.89.

seen, the t-statistic ($t_{\hat{\rho}}$) that tests the null hypothesis that a particular variable has a unit root is small for BR, $\dot{p}$, and NFR, but large for $BR - \dot{p}$ and $NFR - \dot{p}$. On the other hand, the test statistics ($\hat{n}_u$) that tests the null hypothesis that a particular variable is mean stationary is large for BR, $\dot{p}$, and NFR, but small for $BR - \dot{p}$ and $NFR - \dot{p}$. These results indicate that BR, $\dot{p}$, and NFR have a unit root and are thus nonstationary in levels. In contrast $BR - \dot{p}$ and $NFR - \dot{p}$ do not have a unit root and are thus stationary in levels. Table A1 also presents unit roots and mean stationary tests using first differences of BR, $\dot{p}$, and NFR. As can be seen, the test results indicate that first differences of these variables are stationary.

The test for cointegration used is the one proposed in Johansen and Juselius (1990). The test procedure is described in Mehra (1994). Two test statistics—the trace test and the maximum eigenvalue test—are used to evaluate the number of cointegrating relationships. Table A2 presents these two test statistics for determining whether in bivariable systems like $(BR, \dot{p})$, (BR, NFR) and $(NFR, \dot{p})$ there exist a cointegrating vector. Those test results are consistent with the presence of cointegration between variables in each system.

Table A2 Cointegration Test Results

System	Trace Test	Maximum Eigenvalue Test	*k*
$(BR, \dot{p})$	16.2*	11.3	8
(BR, NFR)	36.4*	32.5*	6
$(NFR, \dot{p})$	22.3*	17.1*	8

*Significant at the 5 percent level.

Notes: Trace and maximum eigenvalue tests are tests of the null hypothesis that there is no cointegrating relation in the system. The test used for cointegration is the one proposed on Johansen and Juselius (1990). The lag length in the relevant VAR system is *k* and is chosen using the likelihood ratio test given in Sims (1980). In particular, the VAR model initially was estimated *k* set equal to a maximum number of eight quarters. This unrestricted model was then tested against a restricted model, where *k* is reduced by one, using the likelihood ratio test. The lag length finally selected is the one that results in the rejection of the restricted model.

REFERENCES

Akhtar, M. A. "Monetary Policy and Long-Term Interest Rates: A Survey of Empirical Literature," *Contemporary Economic Policy,* vol. XIII (July 1995), pp. 110–30.

Bernanke, Ben S., and Alan S. Blinder. "The Federal Funds Rate and the Channels of Monetary Transmission," *American Economic Review,* vol. 82 (September 1992), pp. 901–21.

Bernanke, Ben S., and Ilian Mihov. "Measuring Monetary Policy," NBER Working Paper No. 5145. June 1995.

Cook, Timothy, and Thomas Hahn. "The Effect of Changes in the Federal Funds Rate Target on Market Interest Rates in the 1970s," *Journal of Monetary Economics,* vol. 24 (November 1989), pp. 331–51.

Engle, Robert F., David F. Hendry, and Jean-Francois Richard. "Exogeneity," *Econometrica,* vol. 51 (March 1983), pp. 277–304.

Fuhrer, Jeffrey C., and George R. Moore. "Monetary Policy Trade-offs and the Correlation between Nominal Interest Rates and Real Output," *American Economic Review,* vol. 85 (March 1995), pp. 219–39.

Fuller, Wayne A. *Introduction to Statistical Time Series*. New York: Wiley, 1976.

Goodfriend, Marvin. "Interest Rate Policy and the Inflation Scare Problem: 1979–1992," Federal Reserve Bank of Richmond *Economic Quarterly,* vol. 79 (Winter 1993), pp. 1–24.

Hall, A. "Testing for a Unit Root in Time Series with Pretest Data Based Model Selection." Manuscript. North Carolina State University, 1990.

Hoelscher, Gregory. "New Evidence on Deficits and Interest Rates," *Journal of Money, Credit, and Banking,* vol. XVIII (February 1986), pp. 1–17.

Johansen, Soren. "Cointegration in Partial Systems and the Efficiency of Single Equation Analysis," *Journal of Econometrics,* vol. 52 (June 1992), pp. 389–402.

_____, and Katarina Juselius. "Maximum Likelihood Estimation and Inference on Cointegration—With Applications to the Demand for Money," *Oxford Bulletin of Economics and Statistics,* vol. 52 (May 1990), pp. 169–210.

Kwiatkowski, Denis, Peter C.B. Phillips, Peter Schmidt, and Yongcheol Shin. "Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root: How Sure Are We That Economic Time Series Have a Unit Root?" *Journal of Econometrics,* vol. 54 (October–December 1992), pp. 159–78.

Mehra, Yash P. "A Federal Funds Rate Equation," Working Paper 95–3. Federal Reserve Bank of Richmond *Economic Inquiry* (forthcoming 1996).

_____. "An Error-Correction Model of the Long-Term Bond Rate," Federal Reserve Bank of Richmond *Economic Quarterly,* vol. 80 (Fall 1994), pp. 49–68.

Radecki, Lawrence J., and Vincent Reinhart. "The Financial Linkages in the Transmission of Monetary Policy in the United States," in *Bank for International Settlements, National Differences in Interest Rates Transmission*. Basle, Switzerland, 1994.

Roley, V. Vance, and Gordon H. Sellon, Jr. "Monetary Policy Actions and Long-Term Interest Rates," Federal Reserve Bank of Kansas City *Economic Review,* vol. 80 (Fourth Quarter 1995), pp. 73–89.

Sims, Christopher A. "Macroeconomics and Reality," *Econometrica,* vol. 48 (January 1980), pp. 1–48.

Stock, James H., and Mark W. Watson. "A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems," *Econometrica,* vol. 61 (July 1993), pp. 783–820.

Financial Intermediation as Delegated Monitoring: A Simple Example

Douglas W. Diamond

Banks and other financial intermediaries are the main source of external funds to firms. Intermediaries provided more than 50 percent of external funds from 1970 to 1985 in the United States, Japan, the United Kingdom, Germany, and France (Mayer 1990). Why do investors first lend to banks who then lend to borrowers, instead of lending directly? What is the financial technology that gives the banks the ability to serve as middleman? To answer these questions, this article presents a simplified version of the model in Financial Intermediation and Delegated Monitoring (Diamond 1984).[1] The results explain the key role of debt contracts in bank finance and the importance of diversification within financial intermediaries. The framework can be used to understand the organizational form of intermediaries, the role of banks in capital formation, and the effects of policies that limit bank diversification.[2]

Financial intermediaries are agents, or groups of agents, who are delegated the authority to invest in financial assets. In particular, they issue securities in order to buy other securities. A first step in understanding intermediaries is to describe the features of the financial markets where they play an important role and highlight what allows them to provide beneficial services. It is important to understand the financial contracts written by intermediaries, how the contracts differ from those that do not involve an intermediary, and why these are optimal financial contracts. Debt contracts are central to the understanding

The author, the Theodore O. Yntema Professor of Finance at the University of Chicago, Graduate School of Business, is grateful to Tony Kuprianov, Jeff Lacker, and John Weinberg for helpful comments. The views expressed are those of the author and do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] This model is adapted from and extends class notes I have used at the University of Chicago since 1985.

[2] Some models that extend the role of diversification in Diamond (1984) to other interesting issues in financial intermediation are in Ramakrisnan and Thakor (1984), Boyd and Prescott (1986), and Williamson (1987).

of intermediaries. The cost of monitoring and enforcing debt contracts issued directly to investors (widely held debt) is a reason that raising funds through an intermediary can be superior. Debt contracts include contracts issued to intermediaries by the borrowers that they fund (these are bank loans) and the contracts issued by intermediaries when they borrow from investors (these are bank deposits). Portfolio diversification within financial intermediaries is the financial-engineering technology that facilitates a bank's transformation of loans that need costly monitoring and enforcement into bank deposits that do not.

This article both simplifies and extends the analysis in Diamond (1984). Adding an assumption about the probability distribution of the returns of borrowers' projects makes the analysis simpler. There are a few new results that extend the analysis because this article drops the assumption that nonpecuniary penalties can be imposed on borrowers. The change of assumption implies that there is a minimum level of bank profitability required to provide incentives for bankers to properly monitor loans. This article is not a general survey of the financial intermediation literature. Two recent surveys are Hellwig (1991) and Bhattacharya and Thakor (1993). For a survey of the role of debt in corporate finance, see Lacker (1991).

Intermediaries provide services: this is clear because intermediaries issue "secondary" financial assets to buy "primary" financial assets. If an intermediary provided no services, investors who buy the secondary securities issued by the intermediary might as well purchase the primary securities directly and save the intermediary's costs. To explain the sorts of services that intermediaries offer, it is useful to categorize them in terms of a simplified balance sheet. Asset services are those provided to the issuers of the assets held by an intermediary, e.g., to bank borrowers. An intermediary that provides asset services is distinguished by its atypical asset portfolio. Relative to an intermediary that provides no asset services, it will concentrate its portfolio in assets that it has a comparative advantage in holding. The model presented below provides a foundation for understanding this aspect of intermediation, showing that reduced monitoring costs are a source of this comparative advantage.[3] There are other important aspects of intermediation that we do not discuss here: liability services and transformation services. Liability services are those provided to the holder of intermediary liabilities in addition to the services provided by most other securities. Examples include the ability to use bank demand deposits as a means of payment and the personalization of contingent contracts available from life insurance companies. Some liability services, such

[3] Fama (1985) notes that banks issue large certificates of deposit which pay market rates of interest for their risk but are also subject to reserve requirements, implying that the reserve requirements are passed along to borrowers. This is evidence in favor of the idea that banks provide asset services.

as check clearing, are well understood, while others relate to difficult issues in microeconomic theory regarding the role of money. Transformation services involve the conversion of illiquid assets into liquid liabilities, offering improved risk sharing and better liquidity compared with investment in the assets held by intermediaries (see Diamond and Dybvig [1983] and Diamond [1995]). Although there may be interactions between these types of service, this article focuses only on asset services.

If intermediaries provide asset services, they provide services to borrowers who issue assets to them. That is, it matters to the issuer of an asset that the asset is to be held by an intermediary rather than directly by investors. Some costs are lower if the asset is held by an intermediary rather than a large number of individuals. The imperfections that give rise to costs of issuing securities by primary borrowers also give rise to similar costs to an intermediary that issues deposits. I examine how a financial intermediary acting as a middleman can lead to net cost savings, and I develop the implications of this role for the structure of intermediaries. The model yields strong predictions about the contracts used by intermediaries and this provides a setting to analyze important issues in banking policy.

1. AN OVERVIEW OF FINANCIAL INTERMEDIATION: THE COSTS AND BENEFITS OF MONITORING

Theories based on the collection of private information by an intermediary require that there be some benefit to using this additional information in lending. A key result in the agency theory literature is that monitoring by a principal can allow improved contracts. The net demand for this monitoring also depends on the cost of monitoring. This cost depends on the number of lenders who contract with a given borrower.

In contracting situations involving a single lender and a single borrower, one compares the physical cost of monitoring with the resulting savings of contracting costs. Let K be the cost of monitoring and S the savings from monitoring. When there are multiple lenders involved, either each must be able to monitor the additional information directly at a total cost of $m \times K$, where m is the number of lenders per borrower, or the monitoring must be delegated to someone.[4] Delegating the monitoring gives rise to a new private information problem: the person doing the monitoring as agent now has private information. It is not even verifiable whether the monitoring has been undertaken. Delegated monitoring can lead to delegation costs. Let D denote the delegation cost per

[4] Another option is nondelegated monitoring with less duplication of effort, analyzed in Winton (1995). Winton considers multiple prioritized debt contracts, only some of which need monitoring. Because there is still duplicated monitoring, it is qualitatively similar to monitoring by all m investors. To avoid complicating the analysis, this option is not considered here.

borrower. A complete financial intermediary theory based on contracting costs of borrowers must model the delegation costs and explain why intermediation leads to an overall improvement in the set of available contracts. That is, delegated monitoring pays when

$$K + D \leq \min\,[S, m \times K],$$

because $K + D$ is the cost using an intermediary, S is the cost without monitoring, and $m \times K$ is the cost of direct monitoring.

The model in this article illustrates the more general results in Diamond (1984), which analyzes delegation costs by characterizing the organizational structure and contractual form that minimize the costs of delegating monitoring to an intermediary. The first step in studying the benefits of intermediation is to find the best available contracts between borrowers and lenders if there is no intermediary and no monitoring. The optimal unmonitored financial contract between a borrower and lenders is shown to be a debt contract that involves positive expected deadweight liquidation costs which are necessary to provide incentives for repayment.[5] The gross demand for monitoring arises because one can use lower cost contracts (with reduced liquidation costs), if the project's return can be monitored, with an ex ante cost saving of S.

Monitoring is costly, especially if duplicated. If not duplicated, the act of monitoring must be delegated, and then the information obtained is not publicly observed. As a result of the remaining information disadvantage of those who do not monitor, there may be delegation costs associated with providing incentives for delegated monitoring. The best way to delegate monitoring is for the delegated monitor to issue unmonitored debt, which will be subject to liquidation costs. The delegated monitor is a financial intermediary because it borrows from small investors (depositors), using unmonitored debt (deposits) to lend to borrowers (whose loans it monitors).

2. AN EXAMPLE OF OPTIMAL DEBT WITHOUT DELEGATED MONITORING

Consider a borrower who needs to raise a large quantity of capital. All lenders and borrowers are risk neutral, but borrowers have no capital, and each lender's capital to invest is small relative to the amount needed to fund the borrower's investment. The borrower needs to raise 1 (where the units are millions of dollars, and these units will be mentioned only parenthetically), while each investor has $1/m$ units to invest, implying that a borrower needs to raise capital from m investors if $m > 1$. The example assumes that m is very large: $m = 10,000$,

[5] This analysis of optimal debt contracts is extended in Gale and Hellwig (1985). On the value of monitoring in this situation, see Townsend (1979).

and each lender has capital or 0.0001 ($100). Monitoring the borrower costs $K = 0.0002$ ($200), and duplicated monitoring by each of m investors costs $mK = 2$ and is prohibitively expensive. Because monitoring is expensive, one should examine the best contract available without any monitoring.

Investors do not observe the borrower's operations directly, not even its sales or cash flows. How can the lenders write a contract in which they do not need to monitor this information?

The Best Contract without Monitoring

The firm needs to raise 1 ($1 million), and each investor requires an expected return of $r = 5\%$. All lenders and the borrower agree that the borrower has a profitable, positive net present value project to fund, but only the borrower will observe how profitable it turns out to be. The borrower can consume any part of the project's return that he does not pay out to the investor. The interpretation is that the borrower can appropriate the return to himself, since no one else observes the project's success. If the project is a retail store, the borrower can take some sales in cash to himself. More generally, the borrower can inflate costs. In practice, the net cash flows to the firm are very unobservable for many firms. In addition, most other conflicts of interest faced by borrowers can be reinterpreted as equivalent to the borrower's ability to retain underreported cash. The ability to retain underreported cash is simply the most extreme example of a conflict of interest.

The project costs 1 to fund, and its realized value is a random variable with realization denoted by V. The distribution of V, the value of the project, known to all borrowers and lenders is

$$H = 1.4 \text{ million, with probability } P = 0.8,$$

$$L = 1 \text{ million, with probability } 1 - P = 0.2.$$

A Simple Candidate for a Contract is Equity

An equity contract in this context is a profit-sharing agreement, where the profit shared depends on the profits reported by the borrower. Let the fraction of reported profits that goes to the outside investor be a, while the borrower retains a fraction $1 - a$, plus any underreported profits. Suppose that the borrower's contract is just to pay a fraction of reported profits, with no other details or penalties specified. The borrower's payoff, given the true value of V and the reported value, denoted by Z, is $V - aZ$. What profit will the borrower report if he is supposed to pay out a fraction of it? The borrower will choose the smallest value of Z. Supposing that the borrower can't make the lender take a share of a reported loss (by reporting $Z < 0$), the borrower will report $Z = 0$. A simple profit-sharing contract works very poorly when profits cannot be verified. It does not even provide incentives to repay $L = 1$, the minimum possible value

of profit. Even adding the requirement that profit reports can never be less than $L = 1$ does nothing to induce higher payments.

No matter what the true value of V, the best response of the borrower to a profit sharing contract is to pay the lowest possible value. If there is no cost to the borrower of understating the amount, the borrower always does. Even if the lender knows the value of V, if the borrower obtains it first and thus controls it, the lender will not be paid unless the borrower suffers some consequence of not paying.

What Can the Lender Do If the Borrower Claims a Low Amount?

The lender would like to impose a penalty for low payments to give incentives for higher payments. There are two interpretations. The lender can liquidate the project if the borrower pays too little, preventing the borrower from absconding with it, or the lender can impose a nonmonetary penalty on the borrower. Bankruptcy in the world today is some combination of these two actions. In ancient history, the nonmonetary penalties were very common, i.e., debtors' prisons and physical penalties. Such sanctions are now illegal, but the loss of reputation of a borrower of a bankrupt firm is similar to a sanction.

Bankruptcy, Liquidation, and the Optimal Liquidation Policy

Suppose that it is not possible to impose a penalty on the borrower or take other assets (outside the business) that are valued by the borrower. See Diamond (1984) for analysis when these nonpecuniary penalties are possible. The only sanction available to give the borrower an incentive to pay is liquidation of the borrower's assets (as in Diamond [1989, 1991]). To focus on the inefficiency of disrupting firm operations, I assume that liquidating the firm's asset gives no proceeds to the lender or to the borrower. The results are similar when liquidation yields a positive amount that is much less than the value of the unliquidated asset. Liquidation and bankruptcy are useful penalties that a borrower can avoid by paying the debt, but regular liquidation is not a good way to run a firm. How does one specify an optimal financial contract between investor and borrower when one can decide to liquidate (to penalize the borrower) or not, contingent on any payment?

Liquidation is best used as a payment-contingent penalty in the following way. If the lender is ever to liquidate for a given payment, he also should liquidate for all lower payments. Suppose instead that the lender does not liquidate if 1 is paid but will liquidate for some higher payment. Then, whenever the borrower has at least 1, he will avoid liquidation by paying 1 and keep the remainder for himself. This makes meaningless the threat to liquidate given higher payments, because the payment will never exceed 1.

The borrower will pay the lowest amount that avoids liquidation, and keep the rest for himself. The only exception is if the borrower has insufficient funds

to pay that amount. This implies a description of the optimal financial contract without monitoring: select a payment, f, that, if paid, avoids liquidation. The lender then liquidates for all lower payments. This implies that the optimal contract when monitoring is impossible is a debt contract with face f. The face value includes the promised payment of principal and interest.

Determination of the Face Value of Unmonitored Debt

This section determines the minimum face value, f, of unmonitored debt which will lead to payments with an expected return of 5 percent on a loan of 1 ($1 million), or an expected value of 1.05.

Suppose f = 1. When $V = 1$, the borrower pays 1 (paying less would result in liquidation). The borrower gets 0, which is as much as if he paid any lower amount. When $V = 1.4$, the borrower pays 1 (to avoid liquidation), and keeps 0.4 for himself. This implies that with face value of 1, the lender gets 1 for sure, which is less than 1.05 and not acceptable.

Any face value of debt between 1 and 1.4 forces the borrower into liquidation when the project returns 1 but is paid in full when the project returns 1.4. This gives the lender an expected return of $0.8f$, because nothing is received when there is liquidation. Solving for the face value of debt (between 1 and 1.4) that gives lenders a 5 percent expected return solves $0.8f = 1.05$ and yields $f = 1.3125$. Unmonitored debt with that face value works as follows.

Suppose f = 1.3125. When $V = 1$, the borrower pays less than 1.3125, and the asset is liquidated. The borrower gets zero for any payment less than or equal to 1. The best interpretation is that the borrower chooses to pay zero when $V = 1$ because it is the best choice when liquidation is generalized to allow the borrower to keep a positive fraction of the retained cash. This leads the lender to liquidate and receive zero, which occurs with probability 0.2. When $V = 1.4$ the borrower pays 1.3125, avoids liquidation, and keeps $1.4 - 1.3175 = 0.0825$ for himself. This is more than he could get from any smaller payment: any smaller payment gives zero. The payment 1.3125 is received with probability 0.8. Liquidation is only avoided when $V = H$ and the face of 1.3125 is paid. The lender receives 1.3125 with probability 0.8 and zero with probability 0.2, which is an expected payment of $0.8(1.3125) = 1.05$. Any lower face value will give the lender an expected rate of return below 5 percent.

When outside investors cannot observe the cash flows and cannot monitor the business, equity contracts do not work. Enforcing them requires excessively costly monitoring. If this monitoring (sitting on the board of directors or keeping close tabs on the business in other ways) is too costly, then simple financial contracts that do not require monitoring are best. These are debt contracts. They induce the borrower to pay investors because default serves as a penalty that the borrower seeks to avoid.

The analysis can be extended to apply not only to defaults on principal and interest covenants of debt contracts but to any other covenant whose violation implies a potential default on a debt contract. Consider a covenant that might be violated for a variety of hard-to-observe reasons. When it is too costly for lenders to determine the reason for the covenant violation, the covenant will "mean what it says," and involve a default whenever it is violated, rather than being renegotiated based on the reason for the violation.

The Value of Monitoring

Suppose that it is possible for the lender to monitor the value of the borrower's operations. Then, instead of liquidating when less than the face value of debt is paid, the lender who monitors can instead use the threat of liquidation and offer to refrain from liquidation so long as the borrower repays as much as possible. Instead of always or never offering to accept 1 in lieu of liquidation, the lender can offer to accept it when $V = 1$ but not when $V = 1.4$. This policy leads the borrower to pay f when $V = 1.4$ and 1 when $V = 1$. I assume that the lender has all of the bargaining power and will offer to accept less than f only when $V = 1$.

The value of monitoring is the expected savings in financial distress costs, which are equal to $0.2(1) = 0.2$. This is the savings from monitoring, S, described in Section 1. This savings must be compared with the cost of monitoring. The cost of monitoring the value of the borrower's project is K. If there were a single lender, then monitoring would cost K. Duplicated monitoring by each of m lenders would cost mK and would be equivalent to a single lender facing a monitoring cost of mK. I assume that the cost of monitoring is incurred ex ante, before a loan is repaid. Ex ante monitoring implies that the lender must learn in advance about the borrower's business to properly interpret any data about the project's return. In this case, the lender or lenders must establish a costly relationship in order to monitor the borrower. The results can be reinterpreted as also applying to ex post monitoring, where no relationship is needed and where the costs of monitoring are incurred only when the borrower defaults on the debt. If the lender or lenders can commit in advance to monitor if and only if the borrower pays less than face value, the ex ante monitoring results can be adapted as follows. In place of the fixed cost of ex ante monitoring, K, use the expected cost of ex post monitoring, which is the cost K, multiplied by the probability that the borrower must default. If the borrower knows he will be monitored given a default, he will default only when he has no choice, i.e., when $V = 1$ (see Townsend [1979]). The ability to wait to incur the ex post cost of monitoring yields an expected cost of monitoring equivalent to an ex ante cost of monitoring of $(1 - P)K$ or $0.2K$.

3. FINANCIAL INTERMEDIATION

If all m lenders monitor, and m is large, then the cost of monitoring is mK, and monitoring is too expensive. If there were many large investors with personal capital above 1, then monitoring at cost K would be available. With a small supply of large investors who can lend 1 on personal account (fewer such investors than profitable large projects), and no way to delegate monitoring, some projects that would benefit from monitoring will be financed with unmonitored, widely held debt. This section shows how financial intermediaries can be set up to create synthetic large investors. There will be a profit opportunity to set up such intermediaries if none are present. If there are few large investors and no intermediaries, then loans are made at 31.25 percent. Finding a way to make monitored loans at 31.25 percent can allow a banker to make a profit. If intermediation reduces the cost of making monitored loans and there is free entry, bankers will not earn excess profits but instead loan rates will be pushed down.

Suppose that there are no large investors, only small investors each with 0.0001 ($100) to lend, and 10,000 small lenders are needed to finance 1 ($1 million). Suppose the cost of monitoring V is $K = 0.0002$ ($200) for each. If each of 10,000 lenders were to monitor whenever there is a default on the loan, the cost would be 2, which is prohibitive, and no one would monitor. When the monitoring cost is prohibitive, the optimal contract is widely held debt with face value 1.3125 (see the subsection entitled "Determination of the Face Value of Unmonitored Debt"). Delegating monitoring to one agent avoids duplication, but can cause incentive problems for the agent who was delegated the monitoring task. Small lenders will not observe the effort put into monitoring, or the information monitored by the agent. The agent (let's call him or her "the banker") has a conflict of interest with the small lenders. The conflict is similar to the conflict of interest between the borrower and the small lenders. How can the monitoring task be delegated without the need to monitor the monitor? The answer is for the banker to face liquidation as a function of the amount paid to the 10,000 small lenders (depositors). This provides incentives to the banker in the same way it does to a borrower: the banker is always better off paying a sufficient amount to avoid liquidation.

Liquidation is a sanction that the banker tries to avoid. For simplicity and for symmetry with the assumption made about liquidation of borrowers' projects, I assume that liquidation of the bank is only a sanction and yields no cash to the small depositors or to the banker. There are several ways to interpret this high cost of bank liquidation. One interpretation is that when too little is paid to the depositors, the assets of the bank's borrowers are liquidated to make sure that the banker and the borrowers have not colluded to take funds owed to depositors. Another interpretation is that liquidating the bank's assets consumes all of the assets. In addition, because the banker gets zero

when there is a default on deposits, a banker who anticipates that the bank is about to fail will reduce any discretionary component of monitoring. The reduced monitoring will decrease the value of bank assets. The assumption that borrowers and lenders get zero serves as a simple shorthand for these more complicated aspects of the cost of bank liquidation.

Delegated Monitoring without Diversification Does Not Succeed

Suppose that the banker monitors a single loan (runs a one-loan bank) on behalf of the small lenders, and does not diversify across loans. When the borrower's project returns 1, the banker can monitor and collect the 1 without actually liquidating. However, the bank itself would need to be liquidated in this case, because the face value of the bank's debt must exceed 1. If the bank's debt contract with the small depositors has a face value of 1 or less, the small depositors never receive more than 1, which delivers less that the 1.05 expected repayment they need to receive the required 5 percent expected return.[6] If the bank is liquidated when its loan defaults by paying 1, the bank is liquidated whenever the borrower would have been liquidated, had the borrower used widely held debt. Unless the 10,000 lenders each monitor the banker (costing 0.0002 each or a prohibitive total of 2), the one-loan bank will default and be liquidated just as often as the borrower. This one-loan bank example seems to imply that delegating the loan monitoring to the banker will not succeed.

Can the Banker Use Diversification to Reduce Delegation Costs?

Suppose the banker monitored not one loan, but a diversified portfolio of loans. A very simple way to show the value of diversification is to examine the two-loan bank. In particular, suppose the banker monitors the loans of two borrowers whose returns are independently distributed but are otherwise just like that of the single borrower (each loan has a 0.8 probability of returning 1.4 and a 0.2 probability of returning 1). The banker attracts 2 ($2 million) in "deposits" from 20,000 investors and lends it out to two different borrowers. The banker gives each borrower a debt contract with face F ($\$F$ million) and collects F when the borrower has 1.4 and monitors to collect 1 when the borrower has 1. As a result, the banker does not need to use costly liquidation to enforce his loan contract with either borrower. The banker issues unmonitored debt

[6] In the text I ignore the $100 of capital that the banker can contribute, to simplify the explanation. One can slightly lower the face value of debt issued to small outside lenders, but the complication is not very informative. The banker has capital of his or her own to invest. The bank need not raise 1 (million), but only 0.9999 (million). The expected repayment to give a 5 percent expected return is then $(1.05)(0.9999) = 1.04995$. This is equivalent to the case where the banker has none of his own capital but outside investors require a 4.995 percent expected return. The one-loan bank is not viable even when only a 4.995 percent return must be given to outside depositors.

deposits that are widely held, and the bank is liquidated whenever it pays less than face value to any investor. This requires no monitoring by the 20,000 small investors. Let B denote the face value of bank deposits per loan, implying that the two-loan bank has total deposits of $2B$ and each 0.001 ($100) deposit has face value $\frac{1}{10,000}$ B.

Suppose the banker monitors both loans. If both borrowers pay in full, the bank will receive $2F$. If one defaults but not the other, the bank will receive $1+F$. If both default, the bank will receive 1 from each, or 2. The diversification from having two borrowers borrow from the bank will reduce agency costs. The distribution of payments to the bank, if the banker monitors, is as follows:

Payment	Probability	Probability that Payment is $\geq$ this value	Explanation
$2F$	$0.64[P^2]$	0.64	both pay F
$F+1$	$0.32[2(P)(1-P)]$	0.96	one pays F, one 1
2	$0.04[(1-P)^2]$	1.00	both pay 1

Assume that liquidating the bank yields nothing to depositors or to the banker, similar to the liquidation of borrowing firms. The bank has total face value of deposits of $2B$. If the bank must be liquidated when it collects face value of F from one borrower and 1 from the other, it will be liquidated whenever at least one loan defaults, and there will be no possible savings in costs of financial distress. Alternatively, if the bank can and will pay its deposits when one loan defaults, it defaults only when both loans default, and it can reduce the probability of liquidation to $0.04 = (1-P)^2$. To examine when payment of all deposits is possible when just one loan defaults, the total payment received by all depositors will be $2B$ with probability 0.96 and 0 with probability 0.04. The expected payment is $0.96(2)B$. The initial capital needed to make two loans is 2 ($2 million), and it requires a 5 percent expected rate of return, implying that $0.96(2)B = 2(1.05)$, or $2B = 2.1875$, is the promised payment to 2 ($2 million) in deposits. Equivalently, let the promised interest rate on bank deposits be r_B. Then, because $2B = 2(1+r_B)$, the promised interest rate on the bank deposit is $r_B = 9.375\%$.

If the bank is to be able to pay 2.1875 when one loan defaults (paying 1) and the other does not default (paying F), then $1+F$ must be at least 2.1875, and the face value of each loan must satisfy $F \geq 1.1875$. If the bank made loans with this face value, it could avoid liquidation with probability 0.96. In summary, if the bank monitors its loans, it will have the cash and the incentives to pay bank deposits in full with probability 0.96 so long as $F \geq 1.1875$ or the interest rate on bank loans is at least 18.75 percent.

Will the Bank Monitor?

A remaining question is whether the banker will choose to monitor the loans. Without monitoring, the bank would not be able to offer to take 1 when only 1 is available and would instead liquidate the borrower's asset. Monitoring provides no benefit to the banker when all loans pay in full (monitoring is not needed to force a borrower to pay F) nor when all loans default (because the bank fails and is liquidated). The entire increase in the banker's return comes from increasing the return when just one loan defaults.

If the banker who monitors obtains nothing whenever at least one loan defaults, there will be no incentive to monitor. An incentive to monitor requires that monitoring increases the bank's expected payment by at least 0.0002 ($200) per loan. If the banker monitors neither loan, then the bank will fail when just one loan defaults, and the banker will get zero. If a loan that is monitored defaults, and the other loan does not, the banker's return will be $1+F-2B = 1+F-2.1875$. This is the ex post increase in the banker's return due to monitoring. Monitoring one of the loans gives this increased return with the probability that it alone defaults, or with probability 0.16. Monitoring of one of the loans will be in the banker's interest if $0.16(1+F-2.1875)$ exceeds the cost of monitoring or 0.0002. Monitoring one loan will pay if $F \geq 1.18875$. Monitoring both loans gives the same increased return with the probability that one of the two loans is the only default, or with probability 0.32. Monitoring both loans is in the banker's interest so long as $0.32(1+F-2.1875)$ exceeds 0.0004, which also implies $F \geq 1.18875$. So long as the interest rate on bank loans exceeds 18.875 percent, the banker is willing to invest $400 worth of time to monitor all loans because it increases the value of his residual claim on the bank.

The two-loan banker must earn a small profit in excess of the cost of monitoring. The need to provide the bank an incentive to monitor and to avoid bank failure when just one loan defaults (by cross-subsidizing the losses from the defaulting loan with the profit from the nondefaulting loan) leads to profits for the banker who was delegated the monitoring of the loan. The banker will monitor only if it yields a profit, and due to limited liability and limited wealth, the banker never makes deposit payments in excess of loan repayments. The need to provide incentives puts a floor on the banker's expected profit, which is sometimes called a control rent, because the banker's control of decisions requires that the rent (profit) go to him. If further diversification is not possible, either because there are just two loans or because a two-eyed banker can only monitor two loans, bank profits cannot be driven to zero by competition. The two-loan bank has the following profits. The banker gets the residual claim above 2.1875, or

$2.3775 - 2.1875 = 0.19$, with probability 0.64, when neither loan defaults; $2.18875 - 2.1875 = 0.00125$, with probability 0.32, when one loan defaults; and 0, with probability 0.04, when both loans default.

This works out to a total expected payment of 0.122 ($122,000) or $(0.19)0.64 + (0.00125)0.32 = 0.122$. This is a return to the banker of 0.061 per loan, which is in excess of 0.0002 the cost per borrower of monitoring, and the banker earns a control rent of $0.061 - 0.0002 = 0.0608$.[7]

The delegation cost per borrower, D, equals the cost of financial distress of the bank or $0.04(2) = 0.08$, plus control rent to the banker of 0.0608 or a total of 0.1408. All parties are better off with the banker as delegated monitor. The borrower prefers to borrow at 18.875 percent from the bank, versus at 31.25 percent direct. The investors get a 5 percent expected return in either situation. The banker is happy with any claim with an expected payment above $400 and ends up with an expected payment of $122,000.

Summary of Financial Intermediation and Diversification

I consider three types of contracting arrangements: (1) no monitoring: a widely held traded debt contract with face = 1.3125 for each borrower; (2) direct monitoring by investors, which saves distress costs of $S = 0.2$ but costs $mK = 2$; and (3) delegated monitoring by an intermediary, which saves distress costs $S = 0.2$ at cost monitoring plus delegation cost, $K + D = 0.1408$.

Diversification within the intermediary works to make option (3) work by reducing the liquidation cost of providing the bank an incentive to repay small investors. To simplify, I use an example where the diversification from a bank making only two loans was sufficient to give the bank reduced delegation costs. However, it is more generally true that diversification allows financial intermediation to provide low-cost delegated monitoring. The law of large numbers implies that if the bank gets sufficiently diversified across independent loans with expected repayments in excess of the face value of bank deposits, then the chance that it will default on its deposits gets arbitrarily close to zero. In the limit of a perfectly diversified bank, the bank would never default and would face no liquidation costs.[8] In addition, the control rent needed to

[7] One can do a bit better, as in footnote 6. The banker has capital of his or her own to invest. The bank need not raise 2 (million), but only 1.9999 (million). The face value of bank debt owed to depositors is then $1.9999\frac{1.05}{0.96} = 2.18739$. This allows the face value of bank loans to be reduced slightly. The binding constraint is the banker's incentive to monitor, or $0.0004 \leq 0.32(1 + F - 2.18739)$, implying that the face value of bank debt is $F = 1.188614$ (instead of 1.18875). This leads to a payoff to the banker of 0.118721. Because the banker spends 0.0004 of his time on monitoring and is owed 0.0001(1.05) for his 0.0001 capital, there is a total control rent of 0.118216, or 0.059108 per loan.

[8] For a formal limiting argument about well-diversified intermediaries, see Diamond (1984), and for a generalization see Krasa and Villamil (1992).

provide incentives to monitor approaches zero. The delegation cost for the bank approaches zero, and the only cost of intermediation is the (unavoidable) cost of monitoring. Competitive and fully diversified intermediation would drive borrowers' expected cost of capital down to 5.02 percent. In the limit of perfect diversification, the face value of bank debt approaches $F = 1.06275$, which is the solution to $0.8F + 0.2(1) = 1.0502$; it gives the bank a 5 percent expected return after covering the 0.0002 ($200) cost of monitoring. This is too strong because in practice the default risk of borrowers is not independent, it is positively correlated. In addition, the number of loans in the bank's portfolio is limited.

The general message is that diversification allows banks to transform monitored debt into unmonitored debt, delegating the monitoring to bankers. The banks' organizational form minimizes the sum of monitoring and financial distress costs.

Policy Implications

There are important implications of this view of intermediaries. Because there are costs of bank failure, and there are incentive benefits from the bank receiving the profit derived from its monitoring, banks can increase their value by hedging or avoiding risks that they cannot control or reduce via monitoring. For example, monitoring can do nothing to influence the level of riskless interest rates. Thus, there is no incentive reason for the bank to bear general interest rate risk. The bank's high leverage means that a small loss might force a costly default. Hedging of interest rate risk is desirable, through futures markets or interest rate swaps, because it can remove risks that have no incentive value to bank managers. Banks rely on diversification to eliminate the risks of being very highly levered. Unless a risk is intimately related to their monitoring task, banks should avoid risks that are not diversifiable unless the bank can remove the risk from its balance sheet through another (swap or futures) transaction.

Diversification makes bank deposits much safer than bank loans, and in the limit of fully diversified banks with independently distributed loans, bank deposits become riskless. This suggests that even without deposit insurance, deposits ought to be very low risk. Laws that limit bank diversification remove much of the technological advantage of the banking contract. The prohibition on interstate banking in the United States, only recently eliminated, made delegation costs much larger and banks much riskier than they would be without the prohibition. The delegation cost from excessively limited diversification has two components. One is the increased probability of bank failure, which may also have contributed to the historical political pressure for deposit insurance. The other component is excessively high control rents: small undiversified banks require higher levels of future profits to remove their manager's otherwise poor incentives. This suggests that in the United States, where the economy

is large enough to have several competing, well-diversified intermediaries, the increased diversification from geographical deregulation may reduce managerial moral hazard and help eliminate the need for high future bank profits (high charter value) to provide good incentives to bankers. If this is correct, banks and similar financial intermediaries will be more stable in the future than in recent experience in the United States.

4. CONCLUSIONS

The purpose of this article is to clarify the roles of debt and diversification in the financial engineering that is banking. Debt has several roles related to financial intermediation. The right to liquidate on default provides any outside lender with power over the borrower, inducing the borrower to repay the debt. This power is limited by the borrower's right to repay the debt in full and remove the lender's liquidation rights. However, liquidation is potentially inefficient. If the lender cannot monitor the borrower's business, then the lender should liquidate whenever there is a default, no matter what the cause. If the lender can monitor the situation, then the ability to selectively remove the threat to liquidate in return for a concession from the borrower can provide power over the borrower without using inefficient liquidation. Financial intermediaries such as banks can centralize costly monitoring and avoid the duplication of effort of the monitoring of borrowers by small investors. Banks monitor debt (loan) contracts, and issue unmonitored debt (deposit) contracts. Diversification is the financial-engineering technology that makes monitoring of deposit contracts unnecessary when monitoring of loan contracts is necessary. This allows banks to deliver delegated monitoring. Debt, monitoring, and diversification are the keys to understanding the link between financial intermediation and delegated monitoring.

REFERENCES

Bhattacharya, Sudipto, and Anjan V. Thakor. "Contemporary Banking Theory," *Journal of Financial Intermediation*, vol. 3 (October 1993), pp. 2–50.

Boyd, John H., and Edward C. Prescott. "Financial Intermediary-Coalitions," *Journal of Economic Theory*, vol. 38 (April 1986), pp. 211–32.

Diamond, Douglas W. "Liquidity, Banks, and Markets," University of Chicago CRSP Working Paper. December 1995.

_____. "Monitoring and Reputation: The Choice between Bank Loans and Directly Placed Debt," *Journal of Political Economy*, vol. 99 (August 1991), pp. 689–721.

_____. "Reputation Acquisition in Debt Markets," *Journal of Political Economy*, vol. 97 (August 1989), pp. 828–62.

_____. "Financial Intermediation and Delegated Monitoring," *Review of Economic Studies*, vol. 51 (July 1984), pp. 393–414.

_____, and Philip H. Dybvig. "Bank Runs, Deposit Insurance, and Liquidity," *Journal of Political Economy*, vol. 91 (June 1983), pp. 401–19.

Fama, Eugene F. "What's Different about Banks?" *Journal of Monetary Economics*, vol. 15 (January 1985), pp. 29–39.

Gale, Douglas, and Martin Hellwig. "Incentive-Compatible Debt Contracts: The One-Period Problem," *Review of Economic Studies*, vol. 52 (October 1985), pp. 647–64.

Hellwig, Martin. "Banking, Intermediation and Corporate Finance," in Alberto Giovannini and Colin Mayer, eds., *European Financial Integration*. Cambridge: Cambridge University Press, 1991.

Krasa, Stefan, and Anne P. Villamil. "Monitoring the Monitor: An Incentive Structure for a Financial Intermediary," *Journal of Economic Theory*, vol. 57 (June 1992), pp. 197–221.

Lacker, Jeffrey M. "Why Is There Debt?" Federal Reserve Bank of Richmond *Economic Review*, vol. 77 (July/August 1991), pp. 3–19.

Ramakrishnan, Ram T. S., and Anjan V. Thakor. "Information Reliability and a Theory of Financial Intermediation," *Review of Economic Studies*, vol. 51 (July 1984), pp. 415–32.

Townsend, Robert M. "Optimal Contracts and Competitive Markets with Costly State Verification," *Journal of Economic Theory*, vol. 21 (October 1979), pp. 265–93.

Williamson, Stephen D. "Costly Monitoring, Loan Contracts, and Equilibrium Credit Rationing," *Quarterly Journal of Economics*, vol. 102 (February 1987), pp. 135–45.

Winton, Andrew. "Costly State Verification and Multiple Investors: The Role of Seniority," *Review of Financial Studies*, vol. 8 (Spring 1995), pp. 91–123.

A Theory of the Capacity Utilization / Inflation Relationship

Mary G. Finn

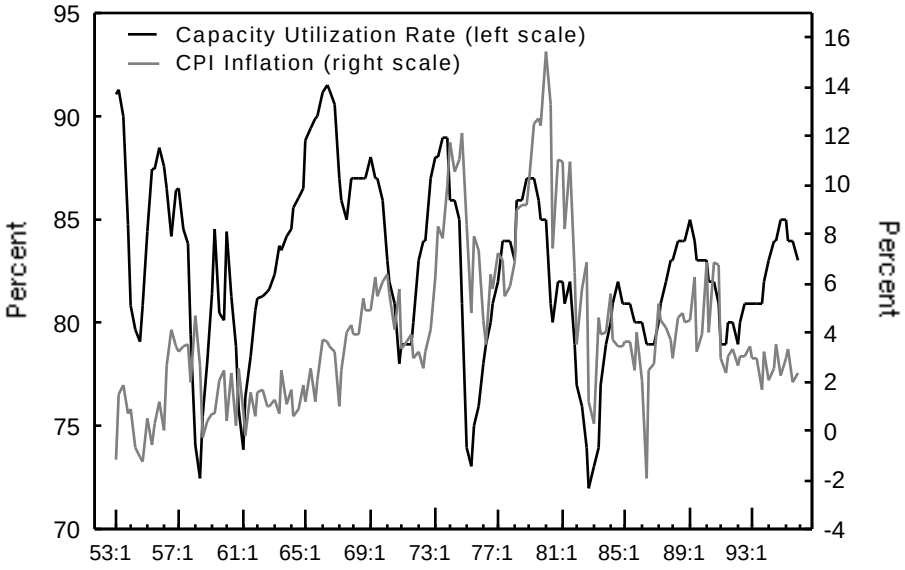
The relationship between capacity utilization and inflation is quite variable. Figure 1 shows the time paths of utilization and inflation for the United States over the period 1953:1 to 1995:4. Two features characterize this relationship. First, inflation and utilization often move in opposite directions. The most dramatic episodes of negative comovement coincided with the 1973/1974 and 1979 periods of sharp energy price rises. Then, utilization plummeted while inflation soared. Second, inflation and utilization also frequently move together. In fact, the instances of positive co-movement slightly dominate those of negative covariation—the average historical correlation between utilization and inflation is 0.09. The question is, why do inflation and utilization behave in this way?

Macroeconomics provides many theories of the relations between real economic activity and inflation. But there is no single theory explaining the foregoing features of the utilization/inflation relationship.[1] Thus to address the question posed above, this article develops a new theory. The new theory is based on the standard neoclassical theory advanced by Kydland and Prescott (1982) and Prescott (1986), which emphasizes the importance of technology shocks for the behavior of real variables such as output, consumption, investment, and employment. Building on the standard theory, the new theory blends together ingredients from various other neoclassical theories. The extensions include endogenous capacity utilization (following Greenwood, Hercowitz, and Huffman [1988]), a role for money and inflation (as in Greenwood and

The author thanks Tom Humphrey, Peter Ireland, Anthony Kuprianov, and Yash Mehra for very helpful comments. The views expressed are those of the author and do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] Finn (1995b) shows that a popular theory, based on a variant of traditional Keynesian theory, does not explain many aspects of the utilization/inflation relationship.

Figure 1 Capacity Utilization Rate and CPI Inflation 1953:1 to 1995:4



Note: CPI inflation is measured quarter to quarter at annualized rates.

Huffman [1987] and Cooley and Hansen [1989]), energy price shocks (see Finn [1995a]), and a rule governing money growth that generally allows the money supply to respond to the state of the economy (following Greenwood and Huffman [1987] and Gavin and Kydland [1995]).

The key ideas of the new theory are as follows. Energy price increases that are as sizeable and surprising as those that occurred in 1973/1974 and 1979, and that are not accompanied by contractions of money growth, cause sharp declines in utilization and rises in inflation. The reason is that a rise in energy prices, by making energy usage more costly, reduces energy input into production. Because the utilization of capital requires energy, utilization must decline along with energy. As productive inputs fall, so too does output. The output contraction induces a rise in inflation, absent an offsetting reduction in money growth. Thus, negative comovement of inflation and utilization occurs in response to energy price shocks.

Exogenous changes in money growth generate a small degree of opposite movement in utilization and inflation. An expansion of money growth directly raises current inflation. By also increasing anticipated future inflation, the rise in money growth reduces the effective return to labor effort. The ensuing reduction of labor implies that the marginal productivity of capital utilization

is now lower and thus causes a fall in utilization. But since the inflation tax on real economic activity is small, the effect of money growth on real variables, including utilization, is small.[2] Therefore, money growth induces a small amount of negative covariation between inflation and utilization.

Allowing money growth to respond significantly and directly to the general state of economic activity, represented by technology, creates a mechanism that results in utilization and inflation moving together whenever technology shocks occur. An increase in technology enhances the productivity of all factors of production, including capital, and thereby stimulates an increase in their usage. The resultant output expansion is a force working to reduce inflation. But when the response of money growth to technology is sufficiently strong, the rise in money growth is the dominating force on inflation and causes inflation to increase. Consequently, accounting for the endogeneity of money growth, technology shocks engender positive comovement of inflation and utilization.

The remainder of the article is organized as follows. Section 1 describes the structure of the model economy and its competitive equilibrium. Section 2 shows how the theory works in principle to explain qualitatively the relationship between utilization and inflation. Section 3 calibrates the model economy to analyze quantitatively the theory's implications for the utilization/inflation relationship. Section 4 offers concluding comments.

1. THE MODEL ECONOMY

This section outlines the model economy's structure and competitive equilibrium.

Structure

The economy produces a good from three factors of production: labor, capital, and energy. In doing so, the degree to which capacity, or the capital stock, is utilized varies endogenously. Capital is never fully utilized because of the costs of utilization, which consist of depreciation and energy costs. The good is consumed, invested, and used to pay for energy purchases from abroad. All households are identical, as are firms. Firms are owned by households. Thus, the economy's representative agent is a combined firm and household. Markets are perfectly competitive and prices are fully flexible.

Money's role is to facilitate transactions. Specifically, money is needed to purchase the consumption good. The supply of money is under the control of a monetary authority. It enters into circulation through transfer payments to the representative agent. Because prices are flexible, money affects real economic activity through one channel only: anticipated future inflation that acts like a

[2] Inflation can affect both the average level and the cyclical behavior of real economic variables. In this study only the cyclical real effects of inflation are analyzed.

tax. By increasing anticipated inflation, a rise in money growth makes activities requiring money, e.g., consumption, more costly relative to activities that do not, e.g., leisure. Thus, increases in money growth cause agents to substitute away from consumption and into leisure.

Stochastic exogenous shocks to production technology, energy prices, and money growth are the sources of fluctuations in the economy. However, not all money growth is exogenous. The monetary authority's rule for money growth allows money growth to partially and directly respond to technology shocks but not to energy price shocks. This endogeneity of money growth captures the idea that the monetary authority accommodates "normal" output fluctuations stemming from technology shocks but not the dramatic fluctuations in output due to the more surprising and larger energy price shocks. A more exact description of the economy's structure follows, with most attention devoted to explaining the extensions on the standard neoclassical model.[3]

The representative agent is infinitely lived with preferences over consumption and labor defined in

$$E_0 \sum_{t=0}^{\infty} \beta^t \, U(c_t, l_t) \quad , \quad U(c_t, l_t) = \log c_t + \eta \, \log(1 - l_t), \tag{1}$$

$$0 < \beta < 1 \quad , \quad \eta > 0,$$

where c and l are the agent's consumption and labor supply, respectively, β is the subjective discount factor, and η is a parameter. Available time each period is normalized at unity. Momentary utility U displays standard features and a unitary elasticity of substitution across consumption and leisure.

The agent produces a good from labor and capital services according to

$$y_t = F(z_t l_t, k_t u_t) = (z_t l_t)^{\alpha} (k_t u_t)^{(1-\alpha)} \quad , \quad 0 < \alpha < 1, \tag{2}$$

where y is the output of the good, z is an exogenous shock to technology, k is the agent's stock of capital in place at the beginning of the period, u is the utilization rate of k, ku is the service flow from capital, and α is labor's output share. The production function F has the usual properties, constant returns to scale and a unitary elasticity of substitution between labor and capital services. It differs from the standard neoclassical production function solely by the inclusion of u. The manner in which u enters into (2) follows Greenwood, Hercowitz, and Huffman (1988), allowing a direct relationship between labor's productivity and utilization.

Goods production also requires energy. In particular, energy compliments capital services in accordance with

$$e_t / k_t = a(u_t) \quad , \quad a(u_t) = \frac{\nu_0}{\nu_1} u_t^{\,\nu 1} \quad , \quad \nu_0 > 0 \quad , \quad \nu_1 > 1, \tag{3}$$

[3] See Hansen (1985) for a description of the standard neoclassical model.

where e is the agent's energy usage and ν_0 and ν_1 are parameters. The technical relationship a in (3) is the same as the one in Finn (1995a). It states that energy is essential to the utilization of capital, with increases in utilization requiring more energy usage per unit of capital, at an increasing rate.

Some of the good is invested by the agent to form capital as follows:

$$k_{t+1} = [1 - \delta(u_t)]k_t + i_t \quad , \quad \delta(u_t) = \frac{\omega_0}{\omega_1} u_t^{\ \omega_1}, \tag{4}$$

$$0 < \delta(\cdot) < 1 \quad , \quad \omega_0 > 0 \quad , \quad \omega_1 > 1,$$

where i is gross investment and ω_0 and ω_1 are parameters. This capital accumulation equation is a standard one except for the variable depreciation rate. Depreciation δ is an increasing convex function of u, as in Greenwood, Hercowitz, and Huffman (1988). Therefore, Keynes's notion of the user cost of capital is captured—higher utilization causes faster depreciation, at an increasing rate, because of wear and tear on the capital stock. In summary, there are two costs of utilization: energy and depreciation, either of which would keep capital from being fully utilized.

At the beginning of any time period, the agent holds money that was carried over from the previous period and receives additional money through a transfer payment from the monetary authority. The agent must use these money balances to purchase the consumption good later in the period. More formally, the agent faces the cash-in-advance constraint:

$$m_{t-1} + (g_t - 1)M_{t-1} \quad \geq \quad P_t c_t \quad , \quad g_t \equiv M_t/M_{t-1}, \tag{5}$$

where m is the agent's chosen money holding, M is the per capita aggregate money supply, the gross growth rate of which is g, $(g_t - 1)M_{t-1}$ is the transfer payment, and P is the price level. The constraint applies to the agent's purchases of the good only when it is used for consumption purposes and not when it is invested or exchanged for energy inputs. Greenwood and Huffman (1987) and Cooley and Hansen (1989) specify a similar transactions role for money.

Another restriction on the agent's activities is the budget constraint, setting total income equal to total spending each period:

$$y_t + [m_{t-1} + (g_t - 1)M_{t-1}]/P_t = c_t + i_t + p_t^e e_t + m_t/P_t, \tag{6}$$

where p^e is the exogenous relative price of energy in terms of the final good. In equation (6), income derives from goods production and total start-of-period money balances; spending is on consumption, investment, energy, and end-of-period money balances. The agent's problem may now be stated: to maximize lifetime utility in (1) subject to the constraints in (2) – (6), taking prices and transfer payments as given.

The description of the economy is completed by specifying the exogenous z and p^e processes and the money authorities' rule determining the evolution

of g. z and p^e are stationary, positively correlated, and independent random variables. Their laws of motion are

$$\log z_t = (1-\rho_z)\log \bar{z} + \rho_z \log z_{t-1} + \epsilon_t^z \quad , \quad 0 < \rho_z < 1, \text{ and} \tag{7}$$

$$\log p_t^e = (1-\rho_p)\log \bar{p}^e + \rho_p \log p_{t-1}^e + \epsilon_t^p \quad , \quad 0 < \rho_p < 1, \tag{8}$$

where ϵ^z and ϵ^p are independent white-noise innovations with zero means and standard deviations σ_z and σ_p, respectively, ρ_z and ρ_p are parameters, and $\bar{z}$ and $\bar{p}^e$ are the respective means of z and p^e. Evidence that technology and the relative price of energy are persistent variables is in Prescott (1986) and Finn (1995a). By treating p^e as exogenous it is implicitly assumed that p^e is determined on a world market that is not substantially affected by the economy under consideration.

The monetary authorities determine money growth in such a way that money growth is a stationary, positively autocorrelated process that partially responds to the state of economic activity. Specifically, the rule governing money growth is

$$\log g_t = \log \bar{g} + \log x_t + \theta \log(z_t/\bar{z}) \quad , \quad \theta > 0, \tag{9}$$

where x is the exogenous component of g with a unitary mean, θ is a parameter, and $\bar{g}$ is the mean of g. The endogenous component of g is $\theta \log(z/\bar{z})$, so called because it depends on the state of the economy as captured by z. Greenwood and Huffman (1987) and Gavin and Kydland (1995) similarly endogenize the money supply process. Thus, the temporary and autocorrelated movements in z around its mean induce the same types of movements in g. The degree of responsiveness of g is directly determined by the size of θ. Accordingly, this monetary rule has the effect of making money growth accommodate the output fluctuations sparked by changes in z but not those engineered by changes in p^e. It is motivated from empirical evidence that money growth positively responds to technology shocks (see Coleman [1996], Gavin and Kydland [1995], and Ireland [1996]). But not all variations in g stem from responses to the economy. An exogenous part of g follows a stationary, positively autocorrelated process that is independent of any other variable:

$$\log x_t = \rho_x \log x_{t-1} + \epsilon_t^x \quad , \quad 0 < \rho_x < 1, \tag{10}$$

where ϵ^x is a white-noise, zero-mean innovation with standard deviation σ_x, which is independent of both ϵ^z and ϵ^p, and ρ_x is a parameter. Thus, purely exogenous and persistent movements in g also occur, consistent with the empirical findings of Cooley and Hansen (1989).

Competitive Equilibrium

Competitive equilibrium is obtained when agents solve their optimization problems and all markets clear. Money market clearing requires $m_t = M_t$. The

competitive equilibrium is determined implicitly by equations (3), (4), (7) – (10), and the following equations:

$$\frac{\eta}{(1-l_t)} = \lambda_t \ \alpha \ \frac{y_t}{l_t}, \tag{11}$$

$$(1-\alpha)\ \frac{y_t}{u_t} \ = \ \omega_0 \ u_t^{(\omega_1-1)} \ k_t + p_t^e \ \nu_0 \ u_t^{(\nu_1-1)} \ k_t, \tag{12}$$

$$\lambda_t = \beta \ E_t\left[\lambda_{t+1}\left\{(1-\alpha)\frac{y_{t+1}}{k_{t+1}} + [1-\delta(u_{t+1})] - p_{t+1}^e a(u_{t+1})\right\}\right], \tag{13}$$

$$y_t = (z_t \ l_t)^{\alpha} \ (k_t \ u_t)^{(1-\alpha)} = c_t + i_t + p_t^e e_t, \tag{14}$$

$$M_t/P_t = c_t, \text{ and} \tag{15}$$

$$\lambda_t = \beta \ E_t\left[\frac{P_t}{P_{t+1} \ c_{t+1}}\right], \tag{16}$$

where λ denotes the marginal utility of real income, i.e., the Lagrange multiplier for the budget constraint (equation [6]). Equation (11) is the intratemporal efficiency condition determining l by equating the marginal utility cost of foregone leisure to the marginal income value of labor's marginal product. The sum of the marginal depreciation and energy costs of utilization is set equal to the marginal product of utilization in equation (12), thereby determining u. Equation (13) is the intertemporal efficiency condition governing investment. It equates the current marginal income cost of investment to the discounted expected future marginal income value of the return to investment. That return is the marginal product of capital plus undepreciated capital less capital's marginal energy cost. The resource constraint for the economy is in equation (14), obtained by imposing the money market clearing condition on (6). The constraint sets net income, $y - p^e e$, equal to expenditure, $c + i$, for the representative agent. Equation (15) states the quantity theory of money, with unitary velocity and consumption as the transaction scale variable.[4] The evolution of money holdings over time is implicitly determined by equation (16). This equation shows that the current marginal real income cost of acquiring one nominal money unit today, λ_t/P_t, equals the discounted expected future marginal consumption value of selling one nominal money unit tomorrow, $\beta \ E_t(1/P_{t+1}c_{t+1})$.

The term $p^e e$ in equation (14) may be interpreted as value added to the production of final output y by the rest-of-the-world's energy good. Thus, $y - p^e e$ is the value added by the domestic economy. In this interpretation, the economy

[4] An implicit assumption is that the interest rate is always positive, ensuring that the cash-in-advance constraint binds each period.

exports final goods to and imports energy goods from the rest of the world. International trade balances each period—the value of exports equals the value of imports, which is $p^e e$.

From equation (16) it follows that anticipated future inflation operates similarly to a tax on economic activity. An increase in future inflation erodes money's expected future purchasing power, causing declines in the marginal utility of real income (see equation [16]) and in desired money holdings. These declines, in turn, induce a reduction in most market activities—such as consumption and labor—stemming from the requirement that money is necessary to finance consumption (see equation [15]).

2. QUALITATIVE WORKINGS OF THE MODEL ECONOMY

To provide some intuition on the workings of the model economy, particularly on the utilization/inflation relationship, this section discusses the main qualitative general equilibrium effects of one-time innovations to each of the three exogenous variables: z, p^e, and x.

Innovation to z

Suppose there is a positive innovation to z, i.e., $\epsilon^z > 0$, causing z to increase. The rise in z has a positive income effect because it improves the relationship between productive inputs and output. In response to the positive income effect, c rises and l falls. By directly increasing labor's marginal productivity, the higher z generates a strong intratemporal substitution force that enhances the rise in c and outweighs the income effect on l, causing l to increase. The higher z also improves the marginal product of u, inducing a rise in u and, concomitantly, in e. As z, l, and u increase, so too does y. Because the expansion of z is persistent, returns to investment are now higher. This rise in returns prompts an intertemporal substitution effect that increases i.

Because the money supply rule directly links g to z, the rise in z unambiguously raises g. What happens to inflation (henceforth denoted by π) depends on the strength of this linkage, i.e., on the size of θ. The reason is that the increases in g and consumption growth exert opposing influences on π; π is increasing in g and decreasing in consumption growth. When the endogenous response of g is significant, i.e., when θ is sufficiently positive, the rise in g exceeds the rise in consumption growth, causing an increase in π. Note that since z is positively autocorrelated and i rises, all of the effects discussed above (relative to the steady state) persist for some time. Therefore, for a sufficiently high value of θ, positive shocks to z induce increases in both u and π. Or, more generally, when monetary policy significantly responds to the state of economic activity represented by z, shocks to z are a source of positive comovement between u and π.

Innovation to p^e

Next consider the effects of an increase in p^e due to a positive realization of ϵ^p. The increase in p^e is tantamount to a terms-of-trade deterioration and, thus, has a negative income effect. As a result of this effect, c falls and l rises. By directly raising the cost of energy, the p^e increase engenders sharp declines in both e and u. Because the contraction of u significantly reduces labor's marginal productivity, a strong intratemporal substitution force is set in motion to reinforce the fall in c and overcome the income effect on l, so that l decreases. The reductions in u and l imply a contraction of y. Since the rise of p^e is persistent, the lower levels of u and l extend into the future. Thus, not only is the future marginal energy cost of capital higher but also the future marginal product of capital is lower. Reduced returns to investment instigate an intertemporal substitution effect that decreases i.

The rule governing money growth ensures g is unaffected by the rise in p^e. Therefore, π unambiguously increases in response to the decline in consumption growth. All of the above effects (relative to the steady state) last into the future because of the positive autocorrelation of p^e and the contraction of i. In short, positive shocks to p^e cause decreases in u and increases in π. More generally, shocks to p^e that are not "offset" by appropriate changes in money growth are a source of opposite movements in u and π.

Innovation to x

Finally, suppose a positive value of ϵ^x occurs, causing a rise in (current) x. The expansion of x directly increases (current) π. Stemming from the serial correlation of the x process, the rise in x generates an increase in expected future x and, thus, in anticipated future π. This signal on future π is important. It is the source of monetary nonneutrality in the model economy. If the signal were absent, the rise in x would simply cause once-and-for-all equiproportionate expansions of the money supply and price level and have no real effects. But when anticipated future π rises, as in the case under discussion, agents expect a shrinkage in the purchasing power of future money balances, which causes a reduction in the marginal utility of real income (λ) and other ensuing real effects.

The fall in λ reduces the marginal income value of the return to work effort, thereby engendering an intratemporal substitution effect that decreases l and c and increases leisure. This fall in c reinforces the rise in π noted above. Because the reduction in l adversely affects the marginal productivity of u, contractions of u and e occur. y must also fall since both u and l are lower. While the effect on capital's future marginal productivity is ambiguous, the current value of λ clearly decreases more than does the future value of λ, because the anticipated inflation effect of the current shock to x diminishes with the passage of time. Therefore, an intertemporal substitution force working through the reduction of the marginal cost relative to the marginal benefit of i is created, which tends to

raise *i*. An alternative way of viewing *i*'s response is to recall that the increase in π erodes money's purchasing power. This erosion makes *c*, which uses money, more costly relative to *i*, which does not use money. Hence, the rise of π induces substitution out of *c* and into *i*. The serially correlated nature of *x* imparts some persistency to all of the effects (relative to the steady state) mentioned above. In summary, the positive shock to *x* sets into motion a decline in *u* and an increase in π. In general, *x* shocks are sources of negative covariation between *u* and π.

3. QUANTITATIVE MODEL ANALYSIS

This section quantitatively explores the model's implications for the relationship between *u* and π.

Methodology

The calibration procedure advanced by Kydland and Prescott (1982) is adopted. In this procedure, values are assigned to the model's parameters and steady-state variables. Some of these values are based on information drawn from other studies or first moments of empirical data. The remaining values are those implied by the model's steady-state relationships. Steady-state variables are denoted using the same notation as before except that time subscripts are omitted. The model's time period is defined as one quarter and the calibration recognizes this definition. Table 1 presents the complete set of calibrated values, with new notation specified in the key. Some details follow.

The values for β, α, δ, and *l* are the same as those often used in quantitative studies (Kydland and Prescott 1991; Greenwood, Hercowitz, and Krusell 1992). $\bar{g}$ equals 1.01, the quarterly average per capita gross growth of M2 in the U.S. economy since 1959 (see Coleman [1996]). The average value of capacity utilization in the United States since 1953 gives 82 percent for *u*. $\bar{p}^e e/y$ is set equal to 0.043, which is the average energy share of output in the U.S. economy (1960–1989) calculated in Finn (1995a). Given the aforementioned number settings, together with the normalization of *y* and $\bar{p}^e$ at unity, the model's steady-state relationships imply numerical solutions for η, ν_0, ν_1, ω_0, ω_1, and all remaining steady-state variables.

No empirical estimate of θ is available in existing studies. Therefore, a sensitivity analysis of θ's values is undertaken here. Specifically, the implications of a range of values for θ from 0 to 0.48, capturing no response to maximum response of *g* to *z*, are analyzed. The upper bound on θ is that value of θ implied by making the variation in the *g* process entirely endogenous or dependent exclusively on the movements in z.[5]

[5] More precisely, setting $\log x_t = 0$ in equation (9) implies $\bar{\theta} = s_g/s_z$, where $\bar{\theta}$ is the upper bound on θ, and s_g and s_z are the respective standard deviations of *g* and *z*. The calibrated values of s_g and s_z follow from the descriptions in the subsequent text and footnote 6.

Table 1 Parameter and Steady-State Variable Values

Preferences	**Other Steady-State Variables**	
$\beta = 0.99$	$y = 1$	
$\eta = 2.07$	$c = 0.774$	
	$i = 0.183$	
Production	$e = 0.043$	$\overline{p}^e e/y = 0.043$
$\alpha = 0.70$	$l = 0.300$	
$\nu_0 = 0.01$	$k = 7.322$	$\delta(u) = 0.025$
$\nu_1 = 1.66$	$u = 0.820$	
$\omega_0 = 0.04$	$\pi = 0.010$	
$\omega_1 = 1.25$		
Monetary Rule		
$\overline{g} = 1.01$		
$\theta \in [0, 0.48]$		
	Stochastic Exogenous Processes	
$\overline{z} = 1.55$	$\rho_z = 0.95$	$\sigma_z = 0.007$
$\overline{p}^e = 1$	$\rho_p = 0.95$	$\sigma_p = 0.032$
	$\rho_x = 0.50$	$\sigma_x = \text{f}(\theta)$, given $s_g = 0.011$

Key: f( ) denotes "function of"; s_g is the standard deviation of g.

Next consider the parameters of the stochastic exogenous processes. The ρ_z, σ_z, and ρ_x values equal those frequently used in other studies (Kydland and Prescott 1991; Cooley and Hansen 1995). The standard deviation of g (denoted by s_g) is set equal to 0.011, the standard deviation of quarterly per capita M2 growth in the United States since 1959 (see Coleman [1996]). The value of σ_x depends on the values of ρ_z, σ_z, ρ_x, s_g, and θ. Thus, the value of σ_x varies as θ changes.[6]

Finn (1995a) estimates the parameters governing the relative price of energy process for the United States (1960–1989). While those estimates do not directly give the values for ρ_p and σ_p of the present study because they pertain to annual data, they do provide some guidance. Consistent with Finn's (1995a) findings of highly persistent energy price movements, ρ_p is equated to 0.95, and of the relative variability of innovations to energy prices and to technology, σ_p equals 0.032.[7]

The quantitative examination of the model focuses on the u, π relationship and consists of two different types of analyses. The first one is an impulse

[6] Equation (7) implies $s_z^2 = \sigma_z^2/(1 - \rho_z^2)$, where s_z is the standard deviation of z. Equation (9) implies $s_x^2 = s_g^2 - \theta^2 s_z^2$, where s_x denotes the standard deviation of x. Equation (10) implies $\sigma_x^2 = (1 - \rho_x^2) s_x^2$. Therefore, σ_x is determined by ρ_z, σ_z, ρ_x, s_g, and θ.

[7] In Finn (1995a), $\sigma_p = 4.57\sigma_z$. Substituting 0.007, the value of σ_z from Table 1, into the latter equation gives $\sigma_p = 0.032$.

response analysis, which traces out the effects of one-time innovations to each of the three exogenous variables z, p^e, and x. The impulse response analysis thus permits isolation of the effects of each shock and does not require knowledge of the shock variances (i.e., σ_z, σ_p, and σ_x). It is the quantitative counterpart of the qualitative discussion in Section 2. The second analysis is a simulation study, where the model economy experiences ongoing innovations to all three exogenous variables. It requires knowledge of the shock variances since these determine the average frequency and/or magnitude of the shocks. The simulation study provides the basis for the computation of the correlation between u and π that summarizes the average relationship between u and π.

Both quantitative exercises require the model's numerical solution for the endogenous variables. The steps involved in the solution are indicated as follows. First, the nonstationary nominal variables are transformed into a stationary form. The transformation divides M_t and P_t by M_{t-1}. Second, the model's parameters and steady-state variables are calibrated. Third, the stationary model is linearized around its steady state and solved using standard solution methods for linear dynamic equations (see Hansen and Sargent [1995]). Fourth, the stationarity-inducing transformation is reversed to give solutions for M_t and P_t.

In addition, for simulation analysis 1,000 random samples of 100 observations on ϵ^z, ϵ^p, and ϵ^x are generated. These samples, together with the model's solution, give rise to 1,000 corresponding samples of 100 observations on the endogenous variables. The correlation between u and π is computed for each sample and then averaged across the 1,000 samples. By averaging across a large number of samples, sampling error is reduced.

Impulse Response Analysis

The impulse response analysis shows the quantitative effects on u and π of once-and-for-all innovations to z, p^e, and x. Specifically, beginning from the steady state (say at time 0), the three experiments are characterized by the time profiles of innovations in the following schematic:

Exogenous Shock To	Time Path of Innovations		
z	$\epsilon_1^z = 0.01$	$\epsilon_t^z = 0$ for $t > 1$	$\epsilon_t^p = \epsilon_t^x = 0$ for all t
p^e	$\epsilon_1^p = 0.50$	$\epsilon_t^p = 0$ for $t > 1$	$\epsilon_t^z = \epsilon_t^x = 0$ for all t
x	$\epsilon_1^x = 0.01$	$\epsilon_t^x = 0$ for $t > 1$	$\epsilon_t^z = \epsilon_t^p = 0$ for all t

The innovations ϵ_1^z and ϵ_1^x are set equal to 1 percent because innovations of that size are sufficient to show the effects of z and x shocks. Moreover, they may be regarded as typical since σ_z and σ_g are close to 1 percent. The case of ϵ_1^p is different. Because of the low energy share of output, a 1 percent shock

to p^e has minuscule economic effects. But, large shocks to p^e have substantive effects. In particular, a 50 percentage point rise in p^e, the approximate value of the p^e increases during the two energy crises of 1973/1974 and 1979 (see Tatom [1991]), significantly affects the economy. Thus, ϵ_1^p is equated to 0.50 to see the effects of one of the largest historical rises in p^e.

A value of θ must be chosen for the z shock experiment only—since in the other two experiments z is held constant and, thus, regardless of θ's value, z does not affect g. As mentioned earlier, a sensitivity analysis of θ was undertaken. It turns out that changes in θ's value within the range [0, 0.48] have only small quantitative effects on real variables, stemming from the fact that the inflation tax on real variables is small. But, the value of θ matters substantially for the behavior of π. When θ is less than 0.25, an increase in z engenders a bigger rise in consumption growth than in g, resulting in a decline in π. For θ greater than (or equal to) 0.25, whenever z rises, the induced expansion of g exceeds that of consumption growth so that π increases. Therefore, recalling the discussion in Section 2, 0.25 is the threshold value of θ at which the endogenous response of g to z becomes sufficiently strong to ensure that z shocks are a source of positive comovement between u and π. The effect of z on π is directly related to the size of θ. In the ensuing z shock experiment, $\theta = 0.35$ is taken as a representative, sufficiently high value of θ.

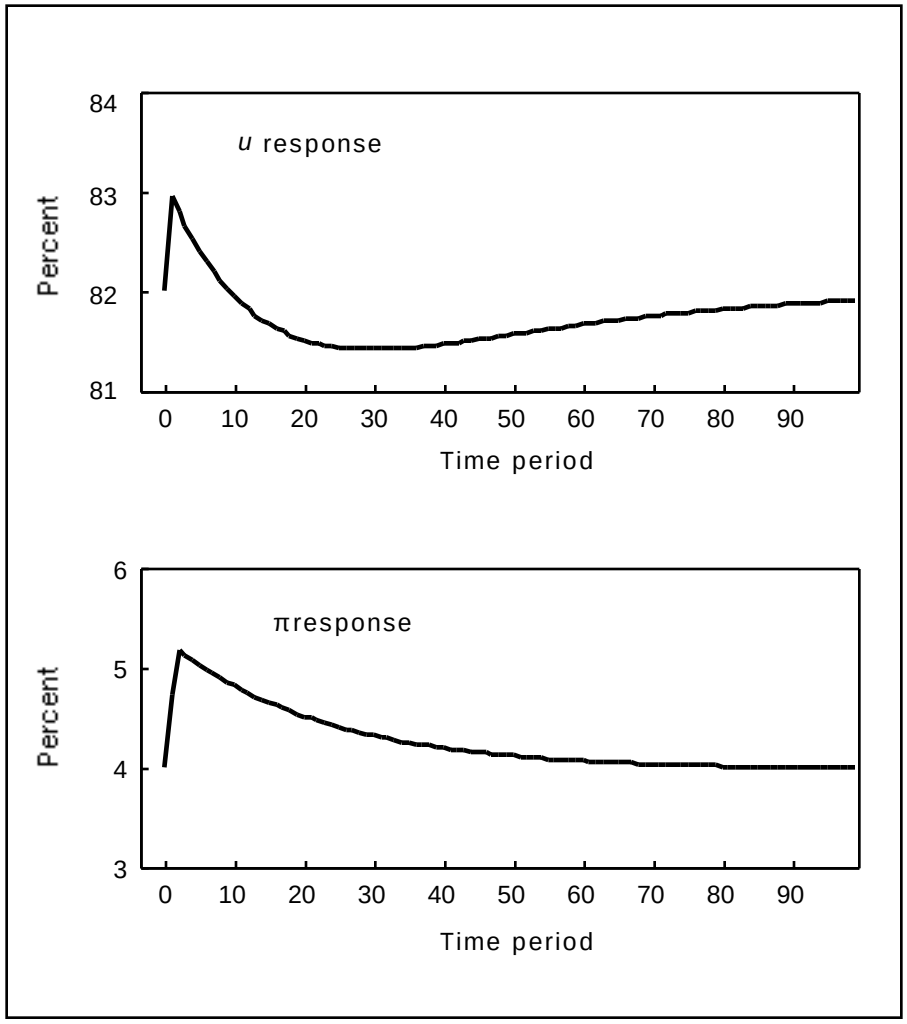
Figure 2 shows the u and π effects of the 1 percent rise in z. At first u rises from 0.82 to 0.83 and then begins to return to its steady-state value.[8] On impact π (expressed at annual rates) increases from 4 percent to 5.2 percent before gradually returning to its steady-state value. Thus, it is seen that z shocks induce strong positive comovement of u and π.

In Figure 3 the responses of u and π to the 50 percent increase in p^e are displayed. u immediately falls from 0.82 to 0.72; subsequently u rises back toward its original value. π jumps from 4 percent to 11.9 percent when the increase in p^e occurs; later π falls to return to its initial value. Consequently, u and π sharply move in different directions when large shocks to p^e occur.

The effects on u and π due to the 1 percent expansion of x are shown in Figure 4. Initially u slightly declines from 0.82 to 0.819 and next rises to return to its steady state. π increases from 4 percent to 9.9 percent at first and subsequently begins its return to the steady state. Therefore, shocks to x cause a small amount of negative covariation between u and π.

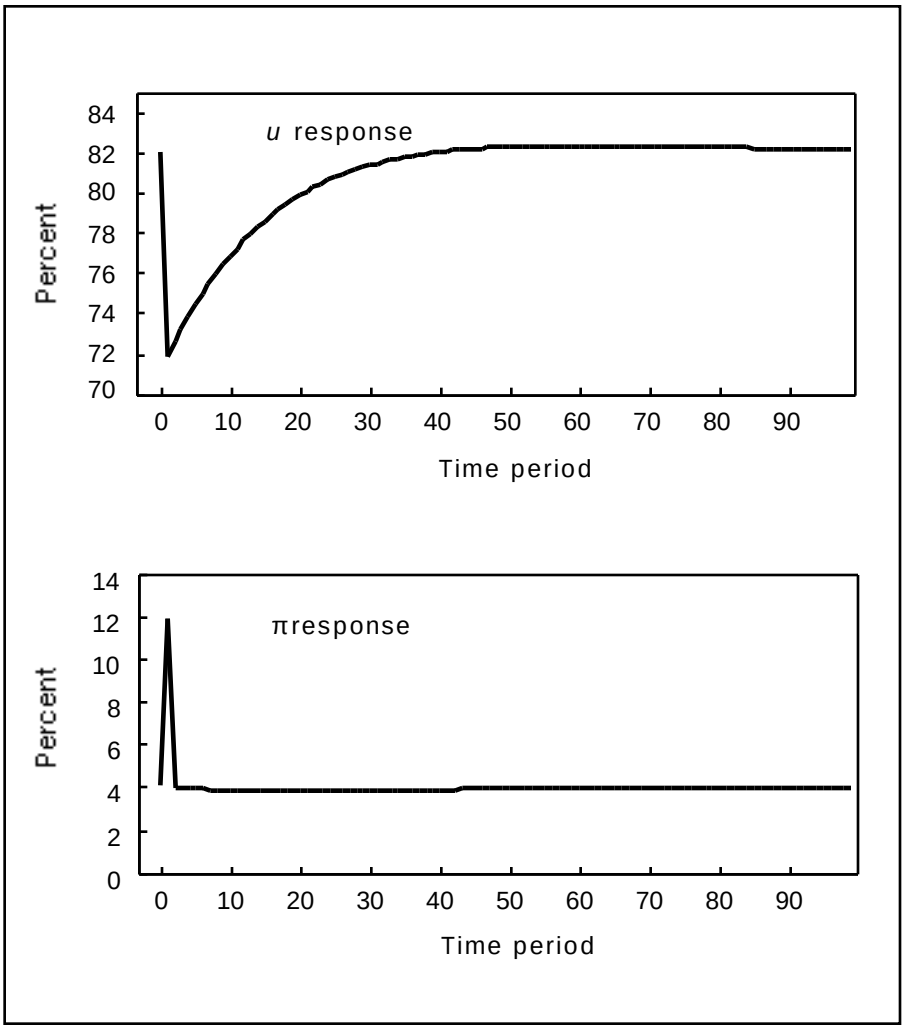
[8] The return path of u is characterized by oscillation. This fluctuation is due to similar behavior in l, which directly affects the marginal productivity of u. The oscillation in l, in turn, stems from the hump-shaped response of k to z shocks, reflecting gradual capital buildup when technology improves, typical in the standard neoclassical model.

Figure 2 Response of u and π to a 1 Percent Rise in z



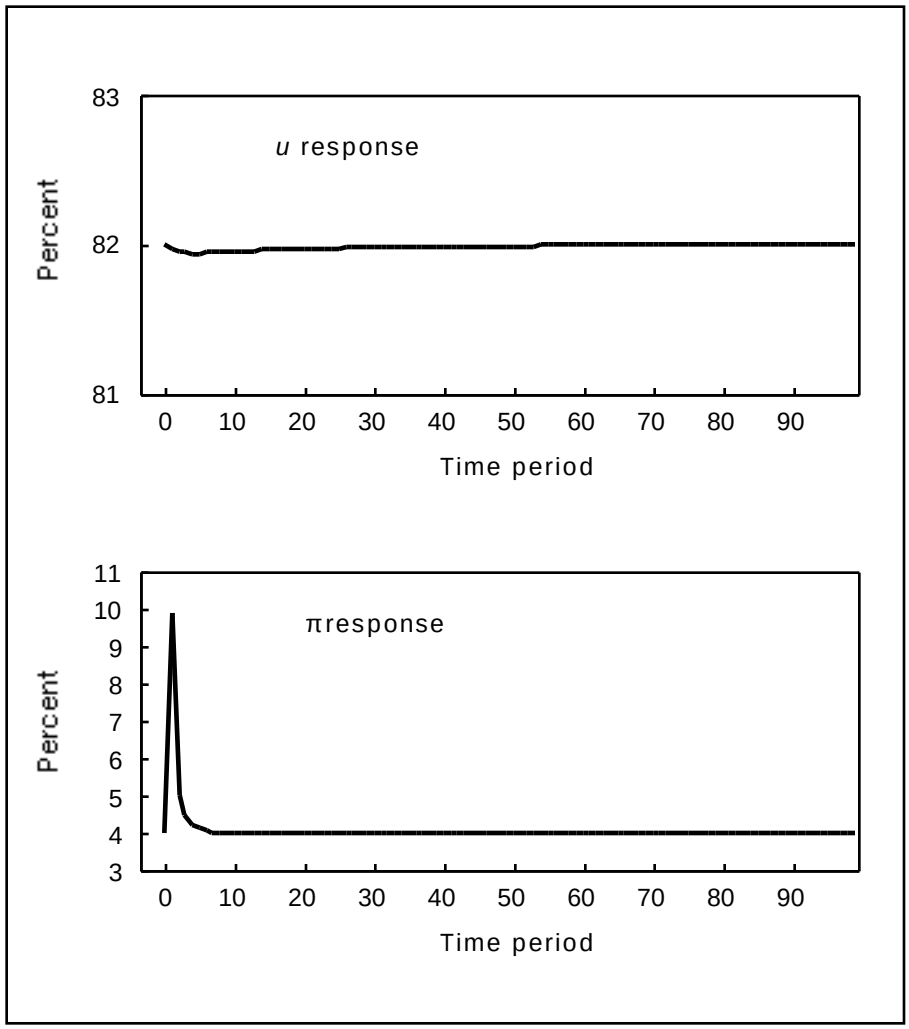
Note: π is expressed at annual percentage rates.

Figure 3 Response of u and π to a 50 Percent Rise in p^e



Note: π is expressed at annual percentage rates.

Figure 4 Response of u and π to a 1 Percent Rise in x



Note: π is expressed at annual percentage rates.

Simulation Results

Table 2 presents the model's correlations between u and π for various values of θ. When g does not respond to z shocks, i.e., when $\theta = 0$, the correlation between u and π is negative. The reason, as explained in more detail before, is in this case all three exogenous shocks cause opposite movements of u and π. But when the endogenous response of g to z is sufficiently strong, specifically when θ is at least 0.25, movements in z give rise to positive comovement of u and π. It turns out that z shocks are so important relative to shocks to p^e and x that for values of θ at least as high as 0.25, the correlation between u and π becomes positive. Moreover, the u, π correlation is increasing in θ because the effect of z on π is directly related to θ.

The model's positive u, π correlations are within close range of the 0.09 value of the correlation between u and π manifest in the U.S. data. Thus, once a significant endogenous response of g to z is accounted for, the model captures quite well the average U.S. historical relationship between u and π.

Table 2 Correlations between u and π

θ	Corr (u, π)
0	−0.11
0.25	0.01
0.30	0.04
0.35	0.06
0.40	0.09
0.48	0.17

Note: Corr ( ) denotes correlation between the variables in parentheses.

4. CONCLUDING COMMENTS

Sometimes in the U.S. economy capacity utilization and inflation move together, a fact that is much emphasized in the popular press. Less noticed is the fact that U.S. capacity utilization and inflation sometimes change in different directions too. Historically, the opposite movements in inflation and utilization have been small in size, with two notable exceptions being the large negative comovements during the energy price crises of 1973/1974 and 1979. On average for the U.S. economy (1953–1995), the instances of positive connections between inflation and utilization slightly dominate those of negative relations because the correlation between inflation and utilization is 0.09. Why do inflation and utilization exhibit such a variable relationship?

This article develops a neoclassical theory to offer an explanation of the utilization/inflation relationship. The causal role of technology shocks, coupled with endogenous monetary responses to economic activity, of energy price

variations, and of changes in money growth are emphasized. The theory shows how technology shocks that are directly accommodated by money growth are an important source of positive comovement between utilization and inflation. On the other hand, according to the theory, substantive shocks to energy prices, in the same order of magnitude as those that occurred in 1973/1974 and 1979, cause dramatically opposite movements in inflation and utilization. Furthermore, the theory explains that changes in money growth cause a small degree of negative covariation of utilization and inflation. The theory's explanation not only works in principle but also meets with *quantitative* success. In particular, it well captures the average correlation between utilization and inflation manifested in the U.S. data.

Because of the neoclassical theory's success in explaining the average utilization/inflation correlation, it would be interesting to use this theory as the basis of further empirical investigations of the utilization/inflation relationship. Specifically, the theory suggests that, underlying the highly variable bivariate relationship between utilization and inflation shown in Figure 1, there is a more stable multivariate empirical relationship between utilization, inflation, technology, energy prices, and money growth. Therefore, working within such a multivariate empirical model might prove useful both in explaining the historical path of inflation and utilization and in forecasting future inflation.

The neoclassical theory developed here incorporates only one source of monetary nonneutrality, the inflation tax. Because of the inflation tax, expansions in money growth cause decreases in utilization while inflation increases. It may be that other channels of monetary nonneutrality, such as sticky prices, are more important for the utilization/inflation relationship because they allow increases in money growth to instead increase *both* utilization and inflation. But the present theory's success in explaining the positive linkages between utilization and inflation *without* such channels creates a strong case that technology shocks and endogenous monetary responses are responsible for much of the utilization/inflation relationship. In so doing, it supports a growing body of theory that stresses the role of technology and endogenous monetary policy in explaining more general relationships between real and nominal economic activity (see, e.g., Gavin and Kydland [1995] and Finn [1996]).

Data Appendix

The data are quarterly and seasonally adjusted for the United States over the period 1953:1 to 1995:4. DRI's database is the source. A detailed description of the data follows.

Capacity Utilization Rate: Total industry (consisting of manufacturing, mining, and utilities) utilization rate for the period 1967:1 to 1995:4. Manufacturing industry utilization rate for the period 1953:1 to 1966:4.

Inflation Rate: CPI annualized quarter-to-quarter inflation.

REFERENCES

Coleman, Wilbur John II. "Money and Output: A Test of Reverse Causation," *American Economic Review,* vol. 86 (March 1996), pp. 90–111.

Cooley, Thomas F., and Gary D. Hansen. "Money and the Business Cycle," in *Frontiers of Business Cycle Research*. Princeton: Princeton University Press, 1995.

_____. "The Inflation Tax in a Real Business Cycle Model," *American Economic Review,* vol. 79 (September 1989), pp. 733–48.

Finn, Mary. "An Equilibrium Theory of Nominal and Real Exchange Rate Comovement." Manuscript. Federal Reserve Bank of Richmond, July 1996.

_____. "Variance Properties of Solow's Productivity Residual and Their Cyclical Implications," *Journal of Economic Dynamics and Control,* vol. 19 (July–September 1995a), pp. 1249–81.

_____. "Is 'High' Capacity Utilization Inflationary?" Federal Reserve Bank of Richmond *Economic Quarterly,* vol. 81 (Winter 1995b), pp. 1–16.

Gavin, William, and Finn Kydland. "Endogenous Money Supply and the Business Cycle," Working Paper 95–010A. St. Louis: Federal Reserve Bank of St. Louis, 1995.

Greenwood, Jeremy, and Gregory W. Huffman. "A Dynamic Equilibrium Model of Inflation and Unemployment," *Journal of Monetary Economics*, vol. 19 (March 1987), pp. 203–28.

Greenwood, Jeremy, Zvi Hercowitz, and Gregory W. Huffman. "Investment, Capacity Utilization, and the Real Business Cycle," *American Economic Review,* vol. 78 (June 1988), pp. 402–17.

Greenwood, Jeremy, Zvi Hercowitz, and Per Krusell. "Macroeconomic Implications of Investment-Specific Technological Change," Federal Reserve Bank of Minneapolis, Institute for Empirical Macroeconomics, vol. 76 (October 1992).

Hansen, Gary D. "Indivisible Labor and the Business Cycle," *Journal of Monetary Economics,* vol. 16 (November 1985), pp. 309–27.

Hansen, Lars, and Thomas Sargent. *Recursive Linear Models of Dynamic Economies*. Chicago: University of Chicago, 1995.

Ireland, Peter. "A Small, Structural, Quarterly Model for Monetary Policy Evaluation." Manuscript. Federal Reserve Bank of Richmond, May 1996.

Kydland, Finn E., and Edward C. Prescott. "Hours and Employment Variation in Business Cycle Theory," *Economic Theory,* vol. 1 (No. 1, 1991), pp. 63–81.

_____. "Time to Build and Aggregate Fluctuations," *Econometrica*, vol. 50 (November 1982), pp. 1345–70.

Prescott, Edward C. "Theory Ahead of Business Cycle Measurement," *Carnegie-Rochester Conference Series on Public Policy*, vol. 25 (Autumn 1986), pp. 11–44.

Tatom, John A. "The 1990 Oil Price Hike in Perspective," Federal Reserve Bank of St. Louis *Review,* vol. 73 (November/December 1991), pp. 3–18.